Title
Computational Protein Design with Multiple Structural and Functional Constraints

Author
Humphris, Elisabeth Lyn

Publication Date
2009

Peer reviewed|Thesis/dissertation

Computational Protein Design with Multiple Functional and Structural Constraints

by

Elisabeth Lyn Humphris

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biophysics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Acknowledgements

I owe the deepest debt of gratitude to Arjun, without whom obtaining my PhD would not have been possible. I would also like to thank my sister for her emotional support, the Kortemme lab members for general discussion and logistical help, and the coffee-shops of the San Francisco Mission for keeping me caffeinated and providing space for me to think and work.

Jim Havranek provided the initial genetic algorithm and multi-specificity code that was modified and used in [1]. Fred Davis helped with generation of the promiscuous protein dataset using the database of protein interfaces (PIBASE) that he developed. I would like to thank the members of my thesis committee for helpful discussions and overall support, Tanja for her unwavering enthusiasm.

Chapters 2 and 3 of this dissertation are revised reprints of manuscripts authored jointly by Humphris and Kortemme appearing in the journals of PLoS Computational Biology [1] and Structure [2] respectively. Tanja Kortemme, the co-author listed on these publications, directed and supervised the research that forms the basis of these chapters.

Chapter 4 describes work that will be submitted for publication.

Abstract

Computational Protein Design with Multiple Functional and Structural Constraints

by

Elisabeth L. Humphris

In this work, a series of computational tools to predict protein sequences compatible with a given three-dimensional protein structure and a set of structural or functional constraints are presented. First, a "multi-constraint" protocol to design protein sequences optimized for multiple criteria is presented. For a number of multi-specific signaling and structural proteins, interface sequences are computationally designed to bind multiple interaction partners and design predictions are compared to naturally occurring amino acid sequences. In many cases, the multi-constraint design algorithm successfully "added up" diverse sequence preferences seen among several characterized binding partners, resulting in the prediction of highly native-like interface sequences. Multi-constraint designed sequences were also found to have overall weaker predicted binding scores than sequences designed to bind only a single interaction partner, suggesting that multi-specificity may come at a cost of affinity. This section concludes by discussing two distinct mechanisms for maintaining multi-specific binding, and providing examples of how the design protocol presented here might be used to rationally design proteins with multiple functional roles.

A method to predict sets of amino acids tolerated at protein-protein interfaces is presented next. By incorporating a flexible backbone move, termed "backrub",

computational predictions of amino acid tolerances at a model interface, human growth hormone with its receptor, are found to closely mimic sequences observed in an experimental phage display dataset. The importance of incorporating backbone flexibility when predicting amino acid tolerance to substitution is discussed and an automated method to computationally predicting sequence libraries to enable challenging protein engineering problems is given.

Finally, a protocol for predicting single amino acid substitutions tolerated for a protein of great biological relevance, HIV-1 protease is presented. In this work multiple constraints present on the HIV-1 protease fold and function are integrated and a reduced set of amino acid mutations (able to be reached by a single mutation at the nucleotide level) was considered. Despite the simplifications inherent in the model, ~80% of amino acid substitutions that occurred in clinical HIV-1 protease sequences were predicted as tolerated. This work further demonstrates that use of a single, fixed backbone as a structural template for design results in overall poorer predictive performance than designing on an ensemble of either crystallographically determined or computationally generated backbone structures.

.

Table of Contents

List of Tables

List of Figures

List of Supplementary Tables

List of Supplementary Figures

Chapter 1

Introduction

This thesis is organized as follows. Chapter 1 gives a brief introduction to proteins, focusing on the relationship between sequence, structure, and function (Section 1.1). An overview of computational protein design, which attempts to predict novel protein sequences compatible with a given protein fold and/or function, follows. Notable computational design successes are highlighted, and possible reasons for computational design to fail to predict folded, functional sequences are discussed in detail (Section 1.2). Two computational tools used throughout this thesis are introduced: the protein sequence design program, RosettaDesign and the "backrub" protocol for computationally generating ensembles of protein backbones (Section 1.3), and an overall outline of the work presented in this thesis is given (Section 1.4).

Chapters 2 and 3 describe the development and testing of two new algorithms that seek to improve upon traditional computational design methodologies. Chapter 4 demonstrates how the novel methodologies introduced in Chapter 2 and 3 can be used to predicting sequence mutations tolerated by a highly biologically relevant protein, HIV-1

protease. Finally, Chapter 5 addresses possible future research directions for the computational protein design methodologies introduced in this work.

1.1. Introduction for the Non-Scientist: The sequence-structure-function paradigm

Proteins are a type of biological molecule, present in all organisms from bacteria to mankind, responsible for an overwhelming number of important biological functions. Some types of proteins, called enzymes, have evolved to catalyze or speed up chemical reactions, breaking down large molecules or building up larger molecules out of smaller components. Other proteins play a structural or mechanical role by binding together to form your hair, fingernails, or cartilage or by providing the contractile force of your muscles. A large number of proteins function by specifically recognizing and binding other partners. For example, the hemoglobin in your blood specifically binds and transports oxygen, proteins in your immune system recognize and bind foreign particles, and a number of proteins are responsible for binding to DNA. The specific recognition and binding of proteins to other molecules or each other can be responsible for setting in motion complex biological signaling cascades, such as an allergy attack or the duplication of an entire cell.

At the simplest level, each protein can be thought of as a linear sequence of "beads" on a string. Each bead represents one of 20 chemically diverse "building blocks", called amino acids, and the total number of beads can vary from ~50 to several thousand. Every amino acid has been assigned a unique single letter (for instance, A for alanine) and thus every protein can be described by writing out an ordered *amino acid*

sequence of single letters. Some proteins have special modifications to amino acids within their sequence, but in most cases the incredible diversity of biological functions observed among naturally occurring proteins is due only to variations in their precise sequence and total number of amino acids.

Most proteins, however, do not perform their biological function as a random coil or a linear chain of amino acids. Chemical properties of amino acids in the linear amino acid sequence interact with each other, forming secondary structures such as flat sheets, three dimensional helices, or relatively unstructured loops. These secondary structures in turn often interact with each other and form a complex three-dimensional shape that brings amino acids far apart in the linear sequence close together in space. The shape that the "string" traces in three-dimensional space is often referred to as the protein *backbone* while the amino acids, or beads, on the backbone are called *side-chains.* It is the three dimensional structure of proteins that mediates all the diverse functional roles described above. For example, cavities or binding surfaces may be created during the folding process through which proteins can interact with chemical substrates, small molecules, or each other.

The relationship between linear amino acid sequence, formation of three-dimensional structure, and biological function is an area currently undergoing intense research. Computation offers great promise to advance understanding of the sequence-structure-function relationship. Even for a relatively small protein of 100 amino acids, the total number of possible linear amino acid sequences available is astronomically large at $(20)^{100}$ or $\sim 10^{130}$. Determining which of these linear combinations of amino acids will adopt a stable, well-folded structure and then experimentally characterizing the resulting

folded three-dimensional coordinates and biological function of such an astronomical number of sequences would be an impossible task.

Computational optimization and search tools offer great promise to efficiently search sequence space and suggest novel folded and functional protein sequence variants. Computation also offers an ideal platform to test our current understanding of the physical forces mediating short- and long- range interactions between amino acids. If accurate models of the physical forces responsible for protein structure and function can be developed, computational tools should be able to make direct predictions of how changes at the amino acid sequence level affect the formation of three-dimensional structure or result in altered biological function.

1.2. Computational Protein Design: Current Strategies and Novel Approaches

In 1961, Anfinsen proposed that the linear amino acid sequences of proteins contain all the information necessary to specify the three dimensional structure of a protein. Today, almost 50 years later, the problem of *protein folding*, or using computation to reliably and accurately predict the precise three dimensional coordinates of a folded, functional protein from knowledge of only its amino acid sequence, remains a formidable computational task. A related, but independent field is that of *computational protein design*. Here, one begins with a known three-dimensional structure of a single protein, or a protein-protein complex, and amino acid sequences compatible with the given starting structure are predicted. *Protein design* has made great progress in dissecting the relationship between altering amino acid sequence and

modifying protein structure and function. Starting from known three-dimensional coordinates, protein design has been able to predict non-native amino acid sequences able to stably adopt the starting fold, and in many cases the sequences are also able to take on altered biological function. Before discussing some of the successes and future challenges of protein design in detail, I first give a basic introduction to some general computational principles.

1.2.1. Principles of Computational Design

Successful protein design typically requires at least three components: a set of starting three-dimensional backbone coordinates, a scoring or energy function which indicates the suitability of each linear amino acid sequence under consideration to adopt the starting three-dimensional structure, and an optimization algorithm to efficiently focus computational time on evaluating those sequences, from among the astoundingly large set of theoretically possible sequences, most likely to score favorably. I will discuss each component in order.

Experimentally, several methods have been developed to visualize the three dimensional structures of proteins, the most common of which determine approximate atomic coordinates of proteins either in solution (using nuclear magnetic resonance, or NMR) or in a crystallized state (using X-rays). There has been exponential growth in the number of protein structures experimentally determined each year since the first protein structure was solved in 1958, and today there are over 60,000 structures (about half of which are independently determined coordinates of the same protein sequence) deposited in the Protein Data Bank (PDB). However, the techniques to experimentally determine protein structures are often difficult and time-consuming, and there are no experimentally

determined structures for many naturally occurring protein sequences. For sequences for which experimental structures do exist, the "fuzziness" (or the structure *resolution*) of the structural picture from which the three-dimension coordinates are determined can vary. The performance of most protein design tools is often best when high-resolution structures whose coordinates have been most accurately determined are used.

Once a three-dimensional structure has been selected, one can computationally remove the identity of any starting side-chains, leaving the atomic coordinates of the backbone fixed. The starting backbone coordinates provide a fixed template onto which differing side-chains sequences can be computationally modeled and scored. However, for any single sequence under consideration, the side-chains can be modeled onto the backbone in any number of possible configurations or orientations. To simplify the possible number of side-chain conformations, protein design tools often consider only a small number of discrete side-chain orientations, where each unique orientation is referred to as a *rotamer*. Standardized libraries of rotamers for all 20 amino acids have been developed to represent the most common side-chain conformations observed in high resolution crystallographic structures.

At each point during a computational design simulation, a unique set of rotamers is selected and scored. (It is the task of the computational search algorithm to ensure rotamers that are likely to be a good fit for the backbone conformation under consideration are ultimately selected for scoring). A variety of scoring functions have been developed, most of which include terms that model physical interactions and forces among the side-chains and backbone atoms, such as attraction and repulsion caused by sterics, electrostatics, solvation, and hydrogen bonding. Often statistical scoring terms

are included to measure whether designed sequences mimic properties that have been commonly observed among experimentally determined structures but as of yet can not be adequately modeled by physical interaction terms. Other statistical terms that have been parameterized to increase performance may also be included to aid in the scoring function being able to correctly discriminate sequences compatible with a given three-dimensional structure from those that are not. Specific examples and details of statistical terms found in the RosettaDesign scoring function are given in Section 1.3.1. In order to increase the speed with which all the terms of the scoring function can be evaluation for each rotameric conformation of side-chains, most scoring functions are limited to including only terms which are pair-wise additive. It should be noted that in the field of computational protein design, scoring functions are often constructed such that sequences with lower (e.g. more negative) scores represent better compatibility for the selected backbone. Details of the scoring function of the computational design program, RosettaDesign (which was used for all computational work in this thesis) can be found in Section 1.3.1.

In Section 1.1, the number of possible sequences for a relatively small, 100-amino acid protein, was calculated to be $\sim 10^{130}$. While the last decade has seen significant advances in the speed of performing computational calculations, sampling and scoring this number of sequences (each of which likely has many distinct possible sets of side-chain conformations), is not yet possible. Computational protein design tools thus often employ any one of a number of optimization protocols, including Monte Carlo simulated annealing, dead-end elimination, and genetic algorithms, to bias sampling of side-chain conformations towards configurations that score better and better with respect to the

starting fixed backbone conformation. In some cases, such as dead end elimination, the search algorithm is guaranteed to output the optimal, or lowest energy, sequence for the given scoring function and starting backbone. Deterministic search algorithms, however, tend to be limited by the size of the design task, such that efficiently redesigning an entire protein sequence may be unfeasible. Other search algorithms, such as the Monte Carlo simulated annealing protocol used by RosettaDesign, are not deterministic and thus not guaranteed to find the optimum possible conformation. Instead, at the end of the design run, the output sequence is a "best guess" at the optimal, or lowest scoring rotameric conformation, possible for the given backbone. In these cases, design simulations may be performed multiple times, with the lowest result seen among all runs taken to be the design output.

In some instances, one may perform protein design for some sequence positions and not others. This might occur, for example, when designing a protein binding surface or interface. Here, one might choose to allow the computational design tools to sample all possible 20 amino acids only at side-chains whose location on the backbone are located within a protein or small molecule binding site. In this case, backbone sites allowed to change their amino acid identities are then referred to *design positions*. To reduce the possibility of physical clashes between sites whose amino acid identities are changing and other sites where amino acid identities remain fixed, side-chain conformation at sites neighboring design positions are often allowed to change rotameric state. These surrounding positions are then referred to as having been *repacked.*

1.2.2. Successes in Computational Design

Protein design methodologies have generally focused on choosing an amino acid sequence optimal for a specific criterion, such as protein stability, interaction energy with a single binding partner, or introduction of catalytic activity into a protein scaffold. Early protein redesign attempts focused on computationally predicting amino acid sequences compatible with protein cores [3] [4]. The first successful redesign of a complete protein was that of a small zinc protein motif [5] and this landmark success was followed by the design of two novel protein folds [6, 7]. These studies demonstrate that computational design search tools and scoring functions have advanced such that prediction of amino acid sequences compatible with adopting a given three-dimensional fold is now possible in many cases.

More recently, there has been significant interest in use of computational tools to design protein sequences with altered functionality. A large number of proteins function through specifically binding and recognizing to each other, and specific protein-protein interactions are known to modulate numerous cellular signaling pathways. Computational design tools able to create novel protein-protein interactions, or interfere with existing protein interactions involved in cellular pathways, offer exciting promise for advancing the creation of biological pharmaceuticals or developing basic science tools for understanding and dissecting complex protein interaction networks.

In general the same physical principles and forces involved in protein folding are also dominant at protein-protein or protein-ligand interfaces. However, deciphering the general principles unpinning the specificity and recognition in binding may pose some unique challenges. The distribution of amino acids in protein interfaces is often

significantly more polar than what is observed in well-packed protein cores. Thus, successful redesign of protein interfaces may, in some instances, require more accurate models of solvation, including the treatment of buried water molecules, electrostatics and hydrogen bonding than needed in the design of protein stability. Additionally, protein binding may be accompanied by conformational changes, which can range from subtle re-arrangements of side chain locations to larger scale movements of loops, helices, or domains. Despite these unique challenges, scoring functions used for predicting sequences stable for a given three-dimensional conformation have been successfully applied to the redesign of protein-protein and protein-ligand interfaces ([8-11], for a review see [12]). Additionally, several groups have attempted to engineer catalytic activity into otherwise non-catalytic protein folds [13, 14], While the catalytic activity of the resulting designed enzymes are still several orders of magnitude less efficient than natural enzymes, these successes indicate that computational tools are advancing such that computer aided design of proteins with sophisticated functionality is now becoming a real possibility.

Despite the computational design accomplishments described above, the rate of success in using computation to design folded, functional proteins with desired functions remains low. This suggests that computational tools may still be far from successfully being able to model all the complex pressures under which naturally occurring sequences have evolved. In the next few sections, several topics that will likely play a large role in the future success of protein design are discussed, including incorporation of multiple functional constraints into protein design, tools to model and incorporate protein

backbone flexibility, and the creation of computational design libraries and tolerance profiles. These topics will be re-visited in greater detail in Chapters 2 and 3.

1.2.3. Designing protein sequences to satisfy multiple functional roles

The field of computational protein design initially began by studying whether one could design, or optimize, a protein sequences for a single fold or function. The examples of protein design given above are primarily examples of *single-state, positive design*. In single-state design, each sequence under consideration is evaluated only with respect to its suitability for fulfilling a single desired property, such as taking on a specific backbone fold, binding a single partner, or catalyzing a single reaction. Whether the same sequences might be able to adopt another backbone fold, bind additional partners, or catalyze multiple reactions is not explicitly considered.

However some design tasks, especially those involving the specificity of binding in protein interfaces, may necessitate *multi-state* computational design methodologies directly able to evaluate the ability of sequences to fulfill multiple properties. A multi-state design protocol, which included explicitly designing against unwanted functional states, was also shown by Havranek and Harbury to be necessary for the specific design of homo- or hetero-dimeric coiled-coil interfaces [15]. Further, Bolon and coworkers found that when traditional single-state positive design was used to create a stable hetero-dimeric protein-protein interface, the alternate homo-dimeric state (which had been ignored during the design process) was almost as stable as the designed hetero-dimeric state [16]. In contrast, when a multi-state *negative design* approach was used, in which both homo-dimeric and hetero-dimeric stability was evaluated for each sequence, the resulting designed sequence assembled almost exclusively into the hetero-dimeric state.

Chapter 2 of this thesis builds on these initial multi-state design successes that designed against an unwanted function by presenting a multi-state positive design methodology to predict protein sequences able to fulfill multiple functional roles. Naturally occurring protein sequences can, in some cases, adjust and adapt their backbone conformations in response to binding, and further, may be responsible for using overlapping interface surfaces to specifically interact with distinct differing binding partners at varying times in cellular signaling pathways. While numerous studies have examined the role particular amino acids may play in ensuring fold stability or binding towards a single partner, less is known about how naturally occurring amino acid sequences have been optimized in order to ensure the correct balance and interplay of multiple properties. The algorithm presented in Chapter 2 is presented as one solution towards incorporating diverse and multiple functional roles into the computational design of protein sequences.

1.2.4. Limitations of Fixed Backbone Design and Methods to Incorporate Conformational Plasticity and Backbone Flexibility

The use of fixed backbone templates, along with libraries of distinct side chain rotameric conformations, has significantly reduced the number of possible conformational states associated with any sequence that must be computationally evaluated and scored. However, naturally occurring protein backbones are often found to adjust their backbone conformation in response to sequence changes [17, 18]. These backbone motions can result in alleviating steric clashes between side-chains that might otherwise be predicted to occur without any backbone movement. Current fixed backbone design methodologies may thus be most likely to provide successful sequence

predictions only in cases where the designed sequence undergoes very little to no backbone adjustments with respect to the starting backbone template. Thus, as it is likely that current fixed backbone methodologies under-predict tolerance to certain sequence changes, the incorporation of backbone flexibility into the protein design process is an area undergoing intense research.

Early incorporations of backbone flexibility relied upon parameterizations of regular secondary structures [19, 20] or random torsional moves [21]. Other groups have sought to incorporate short backbone fragments, obtained from datasets of solved crystallographic structures, with small compensatory torsional moves into protein design [22, 23]. It was with an iterative protocol of incorporating backbone flexibility via fragment insertion and rounds of side-chain design, a novel protein fold, called Top7, was designed [24]. Despite this astounding success of incorporating backbone flexibility into the design of a new protein topology, current flexible backbone methodologies have several limitations. Fragment insertions or small torsional moves can be rejected as incompatible with the starting backbone far more often than they are accepted, and designed backbones can score far worse than the native starting backbone, making discrimination of favorable backbone templates from unfavorable ones more difficult [21, 25].

In Chapter 3, a protocol for using pre-generating ensembles of "near-native" backbones (with root mean squared deviations from the starting fixed backbone structure of typically less than 0.5 Angstroms) in protein design is presented. This protocol independently models sequences of amino acids onto a set of backbone conformations, each of which are a subtle variation of the same fixed backbone, and uses information

gathered from the entire conformational ensemble to determine suitable amino acid substitutions. In Chapter 4, this protocol is modified slightly such that for each sequence change under consideration, the single, most favorable backbone conformation from among the ensemble of conformational variants is chosen.

1.2.5. Using protein design to predict tolerated sequence space and computationally design libraries

Arbitrary protein sequences are rarely functional, as they often fail to take on well-defined three-dimensional structures. In contrast, naturally occurring protein sequences (often referred to as native sequences) can carry out specific functions that are dependent on both their three-dimensional fold and precise amino acid sequence. Changing only a single amino acid within a native sequence (called a *point* mutation*)* can result in a wide range of functional and/or structural effects. Some sequence changes can be thought of as "neutral", if they leave the three-dimensional structure and function of a protein virtually unchanged. Other sequence changes however may result in more drastic effects. These can range from inducing subtle changes in the three-dimensional fold, altering the biological function of the original protein such that it takes on new functionality, or even completely destabilizing the formation of any well-folded, stable three-dimensional structure. Protein design may offer unique insights into discriminating sets of point mutations and amino acid sequence changes the are forbidden (e.g. those causing the protein fold to become unstable) from other sequence changes that may enable the protein to develop new functionality or fold stability. Thus, rather than outputting a single sequence predicted to be optimal for a single design task, computational protein design could also be used more broadly to predict *amino acid*

tolerances at each position in a protein sequence. This would allow the astronomically large theoretical sequence space for a given protein fold to be narrowed down to a more manageable number of sequences, that could then be experimentally screened as combinatorial sequence libraries and tested for novel functionality. Further, use of computationally designed sequence libraries could help overcome some of the limitations in design scoring functions, search methodologies, and fixed backbone approximations discussed above.

Computational design libraries have shown some success when used by several groups to enhance protein folds for added stability and then screen for altered functionality [26, 27]. In Chapter 3, I present a protocol to predict computational design libraries for protein sequences under multiple constraints, using RosettaDesign. This methodology would allow one to directly design libraries enhanced in both fold stability and novel functionality. Further, I present a method to incorporate backbone flexibility, via use of ensemble design, into the computational prediction of design libraries. Current protein design tools are still far from being able to create designer protein sequences at will. Hopefully, the approach presented in Chapter 3 of combining backbone flexibility with experimental screens of computational libraries designed to satisfy multiple criteria will prove useful in enhancing the overall success rate in the design and engineering of protein sequences.

1.3. Introduction to RosettaDesign and the "Backrub" Protocol for Generation Backbone Ensembles

All computational protein design experiments described within this work use RosettaDesign, which has previously been shown to successfully predict binding energy hotspots in protein-protein complexes and has been used to reengineer specificity in protein interfaces [11, 28]. The most general features of the RosettaDesign program, including the sampling methodology and the scoring function, are described in this section. As needed, specific modifications to the standard RosettaDesign protocols of sequence sampling and scoring will occur in the Chapters that follow.

1.3.1. RosettaDesign Scoring and Sampling

The RosettaDesign scoring function is described in detail in [7]. It is dominated by attractive and repulsive Lennard-Jones packing interactions, an orientation-dependent hydrogen term [29], and an implicit solvation model, based on that proposed by Lazaradis-Karplus [30]. Several statistical terms, derived from probabilities of certain features occurring among experimentally determined high-resolution crystallographic structures, are also included. These include backbone dependent terms which account for the suitability of a given amino acid type, or a specific conformation of a given amino acid type, for a given set of backbone phi/psi angles (Ramachandran torsional preferences and rotamer self energies). Also included is a statistical pair term, which encodes the statistical likelihood that two amino acid types are found a certain physical distance from

each other. Unfolded, reference state energy terms are also included for each amino acid type.

A per-residue energy (ERES), consisting of the sum of all scoring terms described above, is calculated for every position among a fixed protein backbone or protein-protein complex, and the sum of all per-residue energies over all protein chains is referred to as the *complex* or *folding* score. In order to facilitate quick estimates of binding energies at protein-protein interfaces, an *interface* score was approximated by summing all the pair-wise terms of the RosettaDesign scoring function between all sites *i* and *j*, where *i* and *j* are on differing protein chains.

A published rotamer library [5] was used for generating the rotameric side-chain conformations of each amino acid residue type, and this library was often further expanded by including the native amino acid PDB conformation as well as additional rotamers around the $\chi 1$ and $\chi 2$ angles. Standard RosettaDesign samples rotamers on a fixed backbone using a Monte-Carlo simulated annealing optimization protocol.

1.3.2. Structure Preparation for RosettaDesign

Structural templates for all RoesttaDesign calculations described in this work were taken directly from the protein data bank (PDB) and, unless otherwise noted, were chosen to have a resolution of 2.5 Angstroms or better. Prior to performing any calculations, each structure was initially prepared by removing all water molecules, heteroatoms, and hydrogen atoms. Hydrogen atoms were then computationally re-introduced as previously described [29]. An initial round of side-chain minimization was performed using the RosettaDesign scoring function and keeping all amino acid identities

and backbone coordinates fixed. After this initial minimization, positions were selected for design and repacking (as described in each Chapter) and all backbone and side chain positions not designated as either to be designed or repacked were kept fixed for all subsequent steps.

1.3.3. Backrub protocol for computationally generating structural ensembles

Chapters 3 and 4 describe research projects that rely on a computational protocol for generating small adjustments in fixed protein backbone conformations, termed "backrub". The backrub move was first presented by Richardson and Richardson [31]to describe subtle variations in backbone conformations observed in high-resolution crystallographic structures. These backbone adjustments were often coupled with the presence of alternative side-chain conformations. A generalized version of the backrub move, as described in [31], was implemented into RosettaDesign by Colin Smith [32, 33]. This protocol for computationally introducing backbone flexibility proved invaluable throughout my thesis work as providing a means to model backbone conformations close in three-dimensional space to a starting PDB protein structure or protein-protein complex of interest. Detailed descriptions of the backrub protocol can be found within the methods sections of Chapters 2 and 3. However, a brief description of how an ensemble of backbones was computationally generated from a single starting crystallographic structure (containing either a single protein backbone or a protein-protein complex) using the backrub move is given here.

Prior to beginning simulations to computationally introduce flexibility, each residue within the starting PDB structure can be designated to either remain fixed or be

designated as allowed to take part in a backrub "move". This allows for generation of flexibility focused around a particular structural section of a starting structure. Residues ending or beginning a stretch of consecutive residues selected to undergo movement are always fixed in this protocol, and thus selecting every residue of a protein to be "backrubbed" will nevertheless result in the first and last amino acids retaining their original conformation. In this thesis, ensembles of backbones with variable conformations were generated from the same starting PDB structure by allowing all residues to take part in the backrub move and then performing 100 or more independent Monte Carlo simulations. Each simulation consisted of randomly selecting two residues that were separated by 1 to 10 intervening residues, rotating the protein backbone segment between the two C_α atoms of the selected residues by 0 to 40 degrees, and optimizing the positions of the C_β and H_α atoms branching off the pivot C_α atoms (according to the CHARMM bond angle potential as described in [33]). This process of selecting and rotating atoms was repeated 10,000 times per Monte Carlo run, and throughout the simulations side chain rotamer moves were interleaved with backbone "backrub" moves.

For the protocols described in this thesis, the lowest energy conformation observed during each of 100 or more Monte Carlo simulations was saved and referred to as a backbone "ensemble member". In some cases (as described in Chapter 4), additional conformational variability was introduced by also saving the last conformation sampled, regardless of its score, during each Monte Carlo simulation or by increasing the Monte Carlo temperature, so that more variable conformations were sampled and accepted. Root mean squared deviations (RMSD) between the lowest energy backbone

conformations generated independently from a single crystal structure by performing the backrub move at low Monte Carlo temperatures (KT=0.6) were usually small, and on the order of 0.3 – 0.4 Angstroms. In contrast, ensembles generated at higher Monte Carlo temperatures (KT=1.2) or including the last conformation sampled rather than only considering the lowest energy conformations, displayed greater conformational variability, and could have average RMSD values of up to 0.6 Angstroms.

1.4. Outline of Thesis

In Chapter 2, I present a computational algorithm designed to optimize protein sequences for multiple functional criteria. The algorithm itself is general in nature and its implementation allows for incorporation of any constraint able to be quantified using an objective function into the optimization procedure. After describing the algorithm in detail, I examine its performance in predicting protein interface sequences able to bind to multiple protein partners. Development and testing of such an algorithm could be a crucial first step towards fulfilling the goal of successfully rationally designing novel proteins able to be expressed and function correctly in a cellular environment and in the context of many possible interaction partners. Finally, I suggest some possible insights into the mechanisms nature might have used to tune protein surfaces to recognize multiple correct partner proteins by comparing computationally predicted amino acid sequences to solutions found in nature.

In Chapter 3, I present a modified version of RosettaDesign able to not just predict the optimal amino acid sequence for a given set of functional constraints, but also able to output a profile of expected amino acid tolerances for each protein design site. I

examine the ability of this modified RosettaDesign tool to predict amino acid sequence tolerance profiles at the interface of a well-studied protein-protein interaction, human growth hormone (hGH) with its receptor (hGHR). Incorporation of backbone flexibility (via the "backrub" move) is shown to be critical for correct discrimination of sequence positions displaying high mutational robustness from those displaying high sensitivity to mutation. It is hoped that the computational prediction of amino acid tolerance profiles, such as those presented in Chapter 3, might prove useful for computationally predicting design libraries that could be experimentally screened for novel proteins variants with new functionality.

In Chapter 4, I combine the multi-constraint design principles tested in Chapter 2 with the ideas of predicting sequence tolerance developed in Chapter 3 in order to make predictions of the total set of single amino acid sequence mutations, reachable by a single mutation at the DNA nucleotide level, tolerated by the biologically important protein, HIV-1 protease. Correct functioning of HIV protease is critical to viral infectivity, and this function is itself constrained by the need of HIV protease to maintain a protein sequence that can stably fold and dimerize, as well as cleave at least 10 endogenous substrates. By computationally including each of these functional constraints into the protein design process, I show that computation can, in many cases, predict which sites in the HIV protease sequence are most mutatable as well as which specific amino acid mutations are most likely to occur in both a neutral (i.e. "drug" free) as well as a selective (i.e. after protease inhibitor treatment) setting. It is hoped that this protocol may be further developed for use as a general tool to preemptively predict mutations likely to

occur in protein sequences in response to complex sets of evolutionary (or drug-induced) pressures.

Chapter 5 describes how the tools presented in Chapters 2 – 4 could be modified and expanded for future use in computational protein design and prediction.

Chapter 2

Designing Sequences to Satisfy Multiple Functional Constraints

This chapter describes a "multi-constraint" protein design protocol to predict sequences optimized for multiple criteria. We use this protocol to examine how naturally occurring proteins are able to maintain specific sets of interactions by characterizing the mechanism and extent to which the interface sequences of 20 multi-specific proteins may be constrained by binding to multiple partners. We find that multi-specific binding can be accommodated by at least 2 distinct patterns. In the simplest case all partners share key interactions, and sequences optimized for binding to either single or multiple partners recover only a subset of native amino acid residues as optimal. More interestingly, for signaling interfaces functioning as network "hubs" we identify a different, "multi-faceted" mode, where each binding partner prefers its own subset of wild-type residues within the promiscuous binding site. Our calculations suggest that these interfaces might have been substantially optimized for multi-specificity. The two strategies make distinct predictions for interface evolution and design. Shared interfaces may be better small

molecule targets, whereas multi-faceted interactions may be more "designable" for altered specificity patterns. While this work has focused on examining multiple protein interactions, the computational methodology we present can easily be generalized for examining how naturally occurring protein sequences have been selected to satisfy a variety of positive and negative constraints, such as taking on multiple, distinct backbone conformations. Finally, we hope that our multi-specific design protocol will provide useful in the future for rationally designing proteins to have desired patterns of altered specificity or fulfill multiple functional roles.

2.1. Introduction

Interactions in protein networks may place constraints on protein interface sequences to maintain correct and avoid unwanted interactions. Proteins have evolved to operate within the context of crowded cellular milieus and complex functional networks [34]. It is not well understood how and to what extent protein sequences and structures are optimized for multiple and likely interdependent properties such as stability and efficiency of folding, low propensity for aggregation, and functional characteristics. Protein-protein interaction networks may impose particular pressures on amino acid residues in protein interfaces if each protein needs to maintain correct and avoid unwanted interactions. Not only specificity of interactions but also a defined level of promiscuity may be required, as it is known that many proteins use regions of overlapping interface residues to bind several partners at different points in time [35].

Protein design predictions offer great promise to help dissect the structural determinants of the interplay between promiscuity and specificity, as well as to create

new molecules that interfere with defined cellular protein-protein interactions with high fidelity and selectivity [36]. Yet in at least some cases sequences completely redesigned on a known protein fold often differ substantially from naturally occurring sequences [5] and the properties of designed proteins can be unusual. Top7, a computationally designed protein with a fold not previously seen in nature [7], has a complex folding landscape strikingly different from that of evolved small, single domain proteins [37]. Thus, if we wish to rationally design new proteins that can be expressed and function correctly in a cellular environment and in the context of many possible interaction partners, it is likely that we will need modeling procedures that are able to consider a variety of requirements defining optimal protein "fitness".

Here we focus on the multiple constraints interaction networks may impose on protein interfaces, both to characterize the evolutionary and biophysical principles shaping these networks, and to develop computational design methods to reengineer them. Previous studies have suggested the importance of negative selection to maintain specificity for a single binding partner [38]. Havranek & Harbury developed a negative design strategy selecting against unwanted partners to predict highly specific coiled-coil interfaces [10]. We extend the idea of incorporating additional selection constraints into computational protein design by examining the inverse problem: how are multiple positive criteria, such as the binding of different partners, accommodated at multi-specific (e.g. promiscuous) protein interfaces?

We perform two computational experiments on 20 multi-specific proteins: First, we optimize each multi-specific interface sequence to maintain interactions with all known structurally characterized partners (multi-constraint protocol). Second, we predict

interface sequences optimal for interacting with each partner individually (single-constraint protocol). We hypothesize that, to the extent that a multi-constraint protocol is a good approximation of pressures acting on promiscuous interfaces, predicted sequences should be more "native-like" when all characterized binding partners are included in the optimization procedure than if only the interaction with a single binding partner is considered. Further, if multiple pressures are playing a role for sequence choices, we can compare the differences in interface sequences selected by each partner alone (single constraint) and all partners considered together (multi-constraint). This comparison should highlight which amino acids are compromises among the various outcomes favored by each binding partner individually.

We show that, overall, inclusion of multiple binding partners during optimization returns sequences closer to those found in native promiscuous interfaces. We find native interface residues predicted to be "hotspots" for each partner remain optimal in the context of optimization for single or multiple partners, while other positions may or may not undergo compromises in order to maintain binding of all partners. These trends resulted in the classification of two broad groups of multi-specific interfaces. In the first group, the number of native residues recovered as optimal was similar for optimizations performed over single or multiple partners. Here key interactions within the interface appeared to be shared, and there was little evidence of compromise in binding preferences among all partners. In contrast, a second group of multi-specific proteins, including signaling "hubs" such as the GTPases, ubiquitin and actin, appeared to have optimized large fractions of their interfaces for binding multiple partners. In these cases, each partner appears to pick and choose subsets of the interface to make key interactions

with, and integrating differences in the binding preferences over all partners often resulted in the native residues being the "optimal compromise" for maintaining binding of all partners.

Our method thus both predicts interface sites responsible for multi-specificity and provides an estimate of the magnitude of pressure exerted on sites by each interaction partner. The method we present here can be used as a predictive tool to study how naturally occurring amino acid sequences might have been constrained by any number of positive or negative criteria - including the ability to adopt two different conformations [39] - or as a protein design tool to rationally redesign variants of native proteins to have a desired set of properties matching user-defined constraints.

2.2. Rationale: Test for optimization and compromise by applying multi- and single-constraint design to promiscuous protein interfaces

We set out to address two main questions (Figure 2.1). First, how optimized are native multi-specific interface sequences for binding multiple partners? It is known that the free energy of binding a single interacting partner is generally not evenly distributed among the native interface residues, but rather some positions are energetically more important than others [40]. Further, phage display experiments have revealed that substantial sequence plasticity may be tolerated at protein interfaces without significantly destabilizing, and often improving, binding of a single partner [41-43].

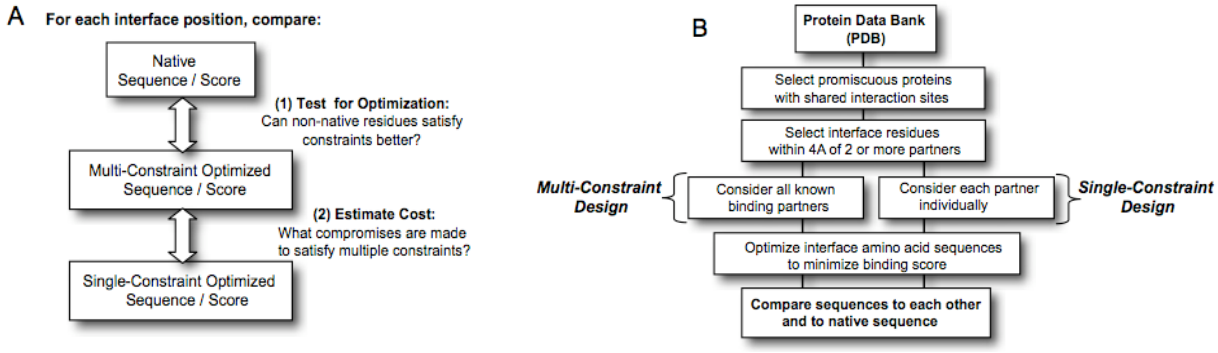


Figure 2.1 Multi-State Design Methodology and Flow Chart
(A) Computational strategy for determining the degree of optimization and predicted cost of multi-specificity. **(B)** Flowchart illustrating the methodology for generating a dataset of multi-specific proteins and computational protocol for predicting sequences optimal for each binding interaction alone (single-constraint) as well as sequences predicted to satisfying binding in the context of all structurally characterized partners (multi-constraint).

Thus, only a subset of a protein sequence may need to be constrained to native in order for a single criterion to be satisfied, while other sequence positions may be less optimized and tolerate a wider set of amino acid types. We hypothesize that the presence of multiple constraints (e.g. multiple binding partners) might substantially restrict interface sequence space such that only native or near native amino acid residues would be tolerated at most sites in multi-specific interfaces. If this is true, sequences that have been computationally designed to optimize binding with all known interaction partners should be more "native-like" than sequences designed to bind each partner independently. Thus, for each promiscuous protein we examine, we compare the sequence predicted to be optimal by our multi-constraint protocol to the wild-type sequence in order to provide an estimate of how extensively each interface is optimized for multi-specificity (*Test for Optimization*, Figure 2.1). Importantly, differences between predicted and wild type sequences could highlight that evolved sequences are not necessarily optimized for maximal affinity but that other pressures are at play.

Secondly, we ask if each binding partner prefers similar interactions throughout the binding site, or if some partners must make energetic compromises in order for multi-specificity to be maintained. In order to address this question, we compare sequences computationally designed to bind only one partner at a time (without consideration of the other characterized partners) with the sequence selected as optimal for interacting with all partners (*Estimate Cost,* Figure 2.1). We reasoned that for a given interface position, if an identical amino acid is chosen when each partner is optimized separately as is selected when all partners are included in the optimization process, key interactions at that site might be highly shared among all partners. In contrast, some single-constraint optimizations might choose an amino acid type different from the one selected in the multi-constraint protocol. For such positions, we can use our scoring function to estimate the degree of compromise occurring between differing preferences seen among the interaction partners.

We imagined two extreme case scenarios: If all binding partners of a given promiscuous protein prefer similar interface sequences ("shared" scenario), single and multi-constraint optimizations are expected to give similar results and comparable agreement with wild-type sequences (termed "native sequence recovery"). If only a core set of shared residues is sufficient for binding to all partners, the total native sequence recovery over the entire interface could be low as the exact amino acid identity of peripheral residues may be less important. Alternatively, each residue in a multi-specific interface could be optimal for only one or few partners ("multi-faceted" scenario). In this case, designed sequences from single constraint simulations would be expected to resemble the wild-type sequence only for certain positions, and these positions could be

different for each partner. The multi-constraint simulations should act to integrate preferences across all partners and would be expected to generate sequences that are more native-like than those resulting from optimization for any single binding partner alone. For this scenario, there could be significant tradeoff between the preferences of differing partners, and amino acid residues within this type of interface could be compromises with respect to the amino acid type preferred by some or all partners. However, for each interface position we hypothesize that there should be an "optimal compromise" that satisfies the constraints imposed by all partners to maintain multi-specific binding.

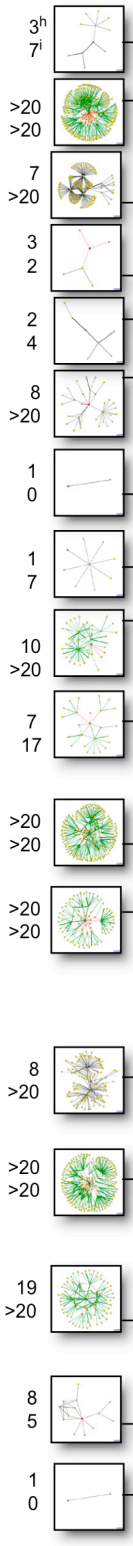
2.3. Computational strategy

Our computational protocol to test for optimization and compromise in multi-specific interfaces outlined above is illustrated schematically in Figure 2.1B. To determine whether the shared or multi-faceted strategies are used in naturally occurring promiscuous interfaces, we first compiled a data set of protein complexes from the PDB (Section 2.4.1). Each promiscuous protein along with all its structurally characterized binding partners is listed in Figure 2.2. In total, we examined 65 PDB complexes, each of which included one of 20 multi-specific proteins. While this analysis is inherently limited by the set of promiscuous proteins characterized in the PDB and ignores much known information on biological interactions, it has the advantage that we can rely on high-resolution structural information for each of the complexes, and hence are more likely to obtain reliable predictions from protein design methods. Our dataset of 20 promiscuous proteins is nevertheless quite broad and includes all SCOP [44] folds (except membrane

proteins) as well as representatives from diverse functional families such as signaling proteins (GTPases, CheY), structural proteins (actin), ubiquitin, and several enzymes (Figure 2.2). Further, in order to estimate the connectivity or number of putative protein-protein interactions for each promiscuous protein in our dataset, we performed a BLAST search against sequences within the database of interacting proteins (DIP, http://dip.doe-mbi.ucla.edu/, [45]). Protein-protein interaction graphs for homologs to the multi-specific proteins in our set (leftmost column of Figure 2.2; e-value $<1*10^{-9}$, Section 2.4.5) suggest at least half of the proteins we analyze can be classified as "hubs" (1^{st} shell nodes > 5; 2^{nd}shell nodes > 15) and many of these proteins are involved in cellular signaling processes.

As we wished to examine promiscuous interface positions believed to be under multiple constraints, only interface positions that had an atom within 4 Å of 2 or more separate binding partners were considered in our analysis. On average, each characterized binding partner contacted 15 (± 4.5) residues in this overlapping set (Figure 2.2). Any conformational changes occurring between the different complexes were taken into account implicitly by using the backbone conformations directly from each complex PDB structure.

Sequence optimizations used a genetic algorithm [10] and fitness for binding was evaluated using inter-molecular scores (Section 2.4.3). Single constraint optimizations minimized the binding score for interaction with a single partner while multiple constraint optimizations minimized the sum of the calculated binding scores over all partners (Section 2.4.3).



Cluster		Promiscuous Protein	Total[a]	Shared[b]	Fold[c]
1		**FYN SH3 Domain**			**all beta**
1[d]	1AVZ[e]	HIV-1 Nef Protein[f]	10	7	(a+b)
2	1M27	SH2 Protein (SAP)	15	7	(a+b)
2		**Importin Beta**			**all alpha**
3	1IBR	Ran	48	10	(a/b)
4	1M5N	Nuclear Localization Signal	36	10	peptide
3		**Ovomucoid Inhibitor**			**small proteins**
5	1HJA	Alpha Chymotrypsin	18	10	all beta
6	1SGP	SGPB	13	10	all beta
4		**Che Y**			**(a/b)**
7	1F4V	FLIM	16	11	(a/b)
8	1FFG	CheA	14	8	(a+b)
9	1KMI	CheZ	9	7	coiled-coil
5		**Thioredoxin**			**(a/b)**
10	1F6M	Thioredoxin Reductase	22	11	(a/b)
11	1T7P	DNA Polymerase	15	11	(a/b)
6		**Phosphocarrier Protein HPR**			**(a+b)**
12	1KKL	HPRK Protein	23	12	(a/b)
13	1RZR	Glucose-resistance regulator	23	12	(a/b)
7		**Interleukin-6 Receptor**			**all beta**
14	1I1R	Viral IL-6	15	14	all alpha
15	1P9M	IL-6[f]	22	14	all alpha
8		**Beta Lactamase**			**alpha beta**
16	1JTD	Inhibitor Protein-II	23	14	mainly beta
17	1JTG	Inhibitor Protein I	34	14	(a+b)
9		**Elastase**			**all beta**
18	1EAI	Elastase Inhibitor	28	15	small protein
19	1FLE	Elafin (Inhibitior)	27	14	small protein
20	1MCV	Hybrid Squash Inhibitor	32	15	small protein
10		**Peroxisome Proliferator Receptor**			**all alpha**
21	1K74	RXR ligand binding domain	26	15	all alpha
22	2PRG	Nuclear Receptor Coactivator SRC-1	25	15	peptide
11		**Ran**			**(a/b)**
23	1A2K	Nuclear Transport Factor II	17	10	(a+b)
24	1I2M	RCC-I	31	11	all beta
25	1IBR	Importin Beta	42	11	all alpha
26	1K5D	Ran Binding Protein I	31	11	(a/b)
27	1WA5	Exportin CSE1P/KAP60P	48	14	all alpha
12		**Ras**			**(a/b)**
28	1BKD	Son of Sevenless-I	38	16	all alpha
29	1HE8	PI-3 Kinase[f]	17	12	(a+b)
30	1K8R	BRY-2RBD[f]	15	10	(a+b)
31	1LFD	Ral-GDS	16	12	(a+b)
32	1WQ1	Ras-GAP	31	17	all alpha
13		**Actin**			**(a/b)**
33	1EQY	Gelsolin	25	11	(a+b)
34	1HLU	Profilin	25	14	(a+b)
35	1LOT	Vitamin D Binding Protein	40	20	all alpha
14		**Transducin Beta Gamma**			**all alpha**
36	1A0R	Phosducin	44	20	(a/b)
37	1GP2	Gi Alpha	34	18	all alpha
38	1OMW	GPCR Kinase II	30	19	(a+b)
15		**1GG1-FC**			**all beta**
39	1ADQ	IGM Rheumatoid Factor[f]	15	14	all beta
40	1DN2	Engineered Peptide	21	20	peptide
41	1FC2	Protein A Fragment	16	11	all alpha
42	1FCC	SGPB Fragment	17	16	(a+b)
16		**Rac**			**(a/b)**
43	1E96	P67 Phox	16	2	all alpha
44	1G4U	GAP SPTP	22	17	all alpha
45	1HE1	Toxin	24	17	all alpha
46	1HH4	Rho GDI	26	13	all beta
47	1I4T	Arfaptin	22	15	all alpha
17		**Ubiquitin**			**(a+b)**
48	1CMX	Modified Ubiquitin	24	16	(a+b)
49	1FXT	Conjugating Enzyme[g]	17	17	(a+b)
50	1NBF	Deubiquitinating Enzyme	35	23	(a+b)
51	1S1Q	TSG-101	16	14	(a+b)
52	1WR6	GGA3-GAT	16	15	all alpha
53	1WRD	TOM-GAT	12	12	all alpha
54	2D3G	HRS-UIM	14	14	all alpha
18		**Cdc42**			**(a/b)**
55	1DOA	Rho GDI	28	16	all beta
56	1GRN	CDC42 GAP	22	13	all alpha
57	1GZS	SOP-E (Toxin)	29	22	all alpha
58	1KI1	Intersectin	28	22	all alpha
59	1NF3	Par	27	11	all beta
19		**RXR Receptor**			**all alpha**
60	1DKF	RAR Receptor	23	21	all alpha
61	1FM6	PPAR Gamma Receptor	26	24	all alpha
62	1MZN	RXR Receptor (Homodimer)	26	23	all alpha
20		**PAPD Chaperone Pilus**			**all beta**
63	1N0L	PAP-E	35	22	all beta
64	1PDK	PAP-K	40	25	all beta
65	1QPP	PAP-D (Homodimer)	29	15	all beta

[a]Total number of interface residues within 4 Angstroms
[b]Number of interface residues in overlapping site
[c]SCOP Fold Classification
[d,e]Interaction partner number/PDB ID
[f]Crystal structure resolution 3 Angstroms or greater
[g]NMR Conformer
[h,i] Predicted number of 1st/2nd node protein-protein interactions

Figure 2.2 Dataset of Multi-Specific Proteins
PDB codes and descriptions of multi-specific proteins and their 65 crystallized interaction partners are given. For each binding partner, the total number of residues it contacts (within 4 Angstroms) on its multi-specific binding protein as well as the number of these residues which are also utilized by at least one other characterized binding partner are given in the ‘‘Total’’ and ‘‘Shared’’ columns. Fold classes are as assigned using SCOP [20]. Protein–protein interaction maps of sequence homologs to the promiscuous proteins in our dataset (see Section 2.4.5) are as taken directly from the Database of Interacting Proteins [21], http://dip.doe-mbi.ucla.edu/, see Supplementary Table A.1). Root nodes are colored red and the number of first (orange) and second (yellow) shell nodes for each map is given on the far left. Edges are color-coded based on the reliability of data used to infer interactions, with green lines indicating data verified by one or more computational methods and red lines depicting unverified high-throughput screens. The width of lines in interaction graphs reflects the number of independent experiments verifying each predicted interaction.

2.4. Methods

2.4.1. Generation of a dataset of multi-specific proteins.

Each domain of every protein-protein interface listed in PIBASE (http://alto.compbio.ucsf.edu/pibase/introduction.html [46]) was classified using the standard SCOP domain definition. SCOP domains were clustered at 90% sequence identity. Clusters containing only intra-protein domain interactions (only one chain in the PDB file) were removed, and clusters with duplicates were merged, leaving 168 clusters. Additional filtering via PDB header descriptions to remove multi-subunit, viral coat, and immunoglobulins/MHC proteins resulted in approximately 50 clusters. All clusters containing multiple structures of the same promiscuous protein interacting with differing binding partners using an overlapping binding site (by visual inspection) were selected for the dataset of multi-specific proteins. Lower resolution structures of redundant protein-protein complexes were discarded, as well as all structures (except 1FXT) determined by NMR. PDB codes of the resulting 20 clusters are given in Figure 2.2.

2.4.2. Energy function and preparation of structures.

All simulations were performed using the RosettaDesign methodology as outlined in [7] and described below. The RosettaDesign scoring function is dominated by attractive and repulsive Lennard-Jones interactions, an orientation-dependent hydrogen bonding term [29], and an implicit solvation model [30]. Side chains from a rotamer library including the native amino acid PDB conformation with additional rotamers around the $\chi 1$ and $\chi 2$ angles [5] were sampled on a fixed backbone using a Monte-Carlo simulated annealing optimization protocol.

All water molecules, heteroatoms, and hydrogens present in the original PDB were removed, and hydrogen atoms were added as previously described [29]. An initial round of side-chain Monte-Carlo minimization was then performed using the RosettaDesign scoring function, keeping all amino acid identities and backbone coordinates fixed, while selecting for the optimal rotamer at each side chain position from the rotamer set as described above. After this initial minimization, all backbone and side chain positions not determined to be in the shared interface were kept fixed for all subsequent steps.

2.4.3. Single and Multi-State Optimization Protocol

Amino acid positions on each promiscuous protein were considered for single and multi-constraint design simulations only if any atom of 2 or more known binding partners was located within 4Å of any atom of the side chain of interest. For promiscuous proteins with 5 or more characterized binding partners, only interface positions with an atom within 4 Å of 3 or more partners were considered. Each single or multi-constraint

optimization allowed all amino acids (except cysteine) to be substituted at each position examined. Positions for which the native residue was a cysteine were disregarded. For all simulations a genetic algorithm was used to generate and propagate putative sequences based on inter-molecular scores, and optimal rotamers for each sequence were chosen separately with consideration of both inter- and intra-molecular interactions by simulated annealing Metropolis Monte Carlo for each fixed backbone as taken from the PDB. This ensured that in the multi-constraint protocol rotameric conformations could differ among binding partners even as identical interface amino acids were scored for each.

Simulations were started with an initial random population of 2000 sequences, and the genetic algorithm was allowed to propagate for 100-200 generations. For single-constraint simulations, fitness was defined to be the inter-molecular score for a single complex while for multi-constraint simulations the fitness was a linear sum of the inter-molecular scores of a given amino acid sequence calculated across all characterized binding partners.

$$\text{FITNESS} = \Sigma\, w_i * (\text{Complex score})$$

For all calculations the weights (w_i) were set uniformly to 1. For single-constraint simulations the sequence that scored optimal with respect to a single complex independently was advanced to the next generation while the multi-constraint protocol advanced the sequence for which the fitness as defined above was minimized. The best (lowest) scoring sequence from each generation was automatically retained and uniform crossover was used to generate the remaining sequences of the population for the following generation. Random mutation of any given interface sequence was allowed for each generation with a probability of 20% at any given interface position. Simulations

converged (dependent on the size of the shared interface) on average within 50-130 generations (see Figure 2.3A and Figure 2.3B).

2.4.4. Per-residue Energetic Analysis.

Over the 20 multi-specific proteins in our dataset, 338 interface residues met the criteria for design. Consideration of each interface position in the context of the 65 characterized binding partners resulted in 1199 individual interactions. For each individual interaction, a per-residue inter-chain score was calculated by summing, for any given residue on chain i, pair-wise contributions to the score from all residues on chain j≠ i. An interface residue was classified as a hotspot for all binding partners for which the per-residue inter-chain score of the original native amino acid in the wild-type complex was calculated to be less than –2 (see pink shading, Supplementary Figure A.4).

Estimates in predicted per-residue improvements (Figure 2.6) in binding affinity were made by calculating, for each binding partner, the difference in per-residue score of the amino acid chosen by single or multi-constraint simulations (Figure 2.6A and Figure 2.6B, respectively) from native. Positions for which the per-residue score for the native amino acid, as well as the amino acid chosen in single- and multi-constraint simulations was zero were eliminated from the analysis. These 214 positions represented cases where one binding partner did not interact with an interface residue in contact with other partners in our dataset (see grey shading, Supplementary Figure A.4)

Estimates of per-residue constraint **(**Figure 2.7) were made by calculating, for each binding partner, the difference in per-residue scores for the amino acid type/rotamer chosen in the single-constraint optimization from the respective score for the amino acid

type/rotamer selected by the multi-constraint protocol. The largest magnitude of difference seen among all partners was the constraint value assigned. For simulations that did not recover the native amino acid type, constraint scores between sequences selected using single and multi-constraint optimization were also calculated and assigned to the native amino acid type.

2.4.5. Generation of protein-protein network graphs.

The complete sequence, as taken from the pdb files, of each promiscuous protein within our dataset was searched against all sequences contained within the Database of Interacting Proteins (http://dip.doe-mbi.ucla.edu/ [45]). Hits were considered as significant if they had an e-value of less than 1*e-9. Protein-protein interaction graphs (Figure 2.2) were shown for sequences predicted to be homologous to Saccharomyces cerevisiae whenever possible. The DIP identification number, organism, e-value, and assigned DIP protein name for interaction graph shown in Figure 2.2 are as given in Supplementary Table A.1.

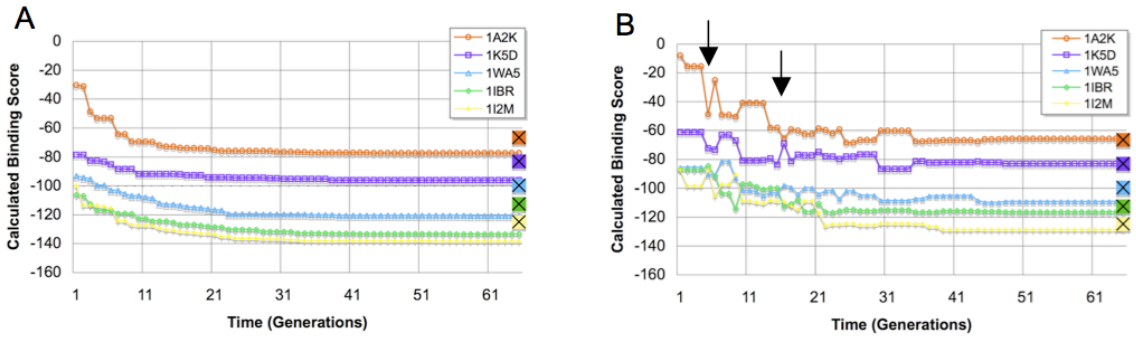
2.4.6. Multiple Sequence Conservation of Ras.

A multiple sequence alignment (MSA) and evolutionary rates for Ras were calculated using the automated web-server http://consurf-hssp.tau.ac.il for the Consurf-HSSP database [47] using the PDB ID code 1WQ1. Evolutionary conservation scores (1-10, 10 most conserved) were 9 and 8 for 32Y and 67M respectively. 90% (186/206) of sequences within the MSA for the native position 32Y contained either a Y or H while

89% (184/206) of sequences at the native position 67M contained (H, I, L, M, Q, V). Multi-constraint simulations selected 32H and 67H as optimal, respectively.

2.5. An example case study: Ran GTPase shows multi-faceted binding

Before discussing results over the entire dataset (complete data for all promiscuous proteins in our set are available as Supplementary Figure A.4), we consider as a representative example the promiscuous protein Ran with 5 of its structurally characterized interaction partners (Figure 2.3). One multi-constraint and five single-constraint optimizations were performed for the Ran set. The trajectories of the five independent single-constraint optimizations monotonously decrease in score at each generation, and in each case the converged final sequence is predicted to have a binding score better than wild type (Figure 2.3A, squares at final generation). Additionally, the sequences selected as optimal in each single constraint simulation differed significantly from native (22-39% native sequence recovery, plus signs in Figure 2.3C).



C

	Ran Interface Residues																		Identical to Native	
DESIGN PROTOCOL	43	44	45	67	68	70	71	74	75	76	77	78	79	95	96	109	110	134	#	%
NATIVE	L	G	V	A	G	E	K	G	L	R	D	G	Y	R	V	R	R	K	18	
MULTI	E	+	M	G	+	H	+	+	+	+	+	+	A	+	H	+	+	+	12	67%
SINGLE 1A2K	Q	+	W	V	+	W	+	H	G	+	R	+	Q	L	R	Y	M	V	5	28%
1I2M	R	R	E	H	+	R	E	+	G	+	R	L	V	+	H	+	F	+	6	33%
1IBR	K	M	L	T	+	W	W	S	+	+	+	T	F	N	G	Q	+	W	5	28%
1K5D	+	+	P	G	F	Y	R	D	+	L	I	F	G	+	R	G	Y	R	4	22%
1WA5	E	S	E	T	Y	R	D	+	+	+	+	M	E	+	K	+	E	+	7	39%

Figure 2.3 Simulation Trajectories and Designed Sequences for the Multi-Specific Protein Ran
Trajectories of single-constraint and multi-constraint optimizations are shown in (A) and (B). PDB codes of complexes used to model the five different Ran binding partners are given in the legend. For reference, the score of the native amino acid sequence for each binding partner is marked on the y-axis (squares, final generation). Scores among partners are correlated for multi-constraint simulations (arrows). Optimal interface sequences taken from the endpoint of the trajectories in (A) and (B) are shown in (C). The first row in the table contains the interface residue PDB numbering, the second row lists the native sequence (red), and the following rows list sequences predicted to be optimal in each simulation: multi- constraint (second sequence), single-constraint (third through seventh sequences). Plus signs in the table denote that the wild-type amino acid residue type was recovered as optimal. The number and percent of interface residues recovered as identical to native is shown for each simulation in the rightmost column. Grey shading denotes interface positions not within 4 Angstroms of the shaded interaction partner (see Section 2.4.3).

In contrast, the trajectories of the multi-constraint simulation show correlated changes in binding scores as each sampled sequence is evaluated separately in the context of the 5 complexes (Figure 2.3B). Cases where the simulation makes tradeoffs that are more favorable to some partners and less favorable for others can be clearly seen (arrows in Figure 2.3A). Here, the sum of scores over all complexes decreases with time and the final converged sequence ranks closer to the native score than the sequences selected by the single partner optimizations (compare endpoints of trajectories of Figure 2.3A and Figure 2.3B with squares). Most notably, the amino acid sequence selected as optimal by the multi-constraint protocol is quite similar to the evolved wild-type sequence (67% identical to wild type, plus signs in Figure 2.3C).

In the Ran example, the high native sequence recovery seen in the multi-constraint optimization indicates that a significant fraction of wild type residues in this promiscuous interface is optimized for multi-specificity by “adding up” information from single partner optimizations **(**Figure 2.3C). This is consistent with the multi-faceted

scenario described above. Further, the multi-constraint trajectories illustrate that there may be tradeoffs in preferences among the binding partners (Figure 2.3B, arrows), and comparison of sequences selected by the single and multi-constraint simulations suggest interface positions where the wild-type residue may represent a compromise to allow promiscuity.

The top row of Figure 2.4 (panels A-F) shows such instance where several single-constraint optimizations select residues differing from native, yet the multi-constraint optimization integrates the single partner preferences to recover the wild type glycine (single constraint models shown for Figure 2.3C, 1st box, residue 74). The design simulations predict three of Ran's binding partners (1A2K.pdb, 1IBR.pdb, 1K5D. Figure 2.4A,C-D) prefer side-chains larger that the wild-type glycine that have additional side-chain hydrogen bonding capability. However tight steric constraints for binding the remaining two partners (1I2M.pdb and 1WA5.pdb, Figure 2.4B,E) necessitate glycine to be the "optimal" compromise for this interface position. Similar instances of compromise at interface positions that are under substantial steric constraint with a subset of the interaction partners are a common pattern in our dataset; many of these cases involve wild-type glycine residues.

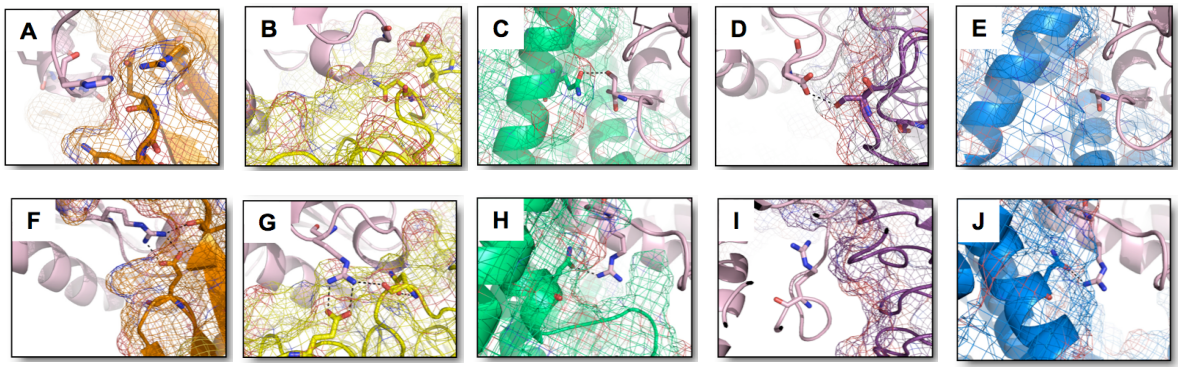


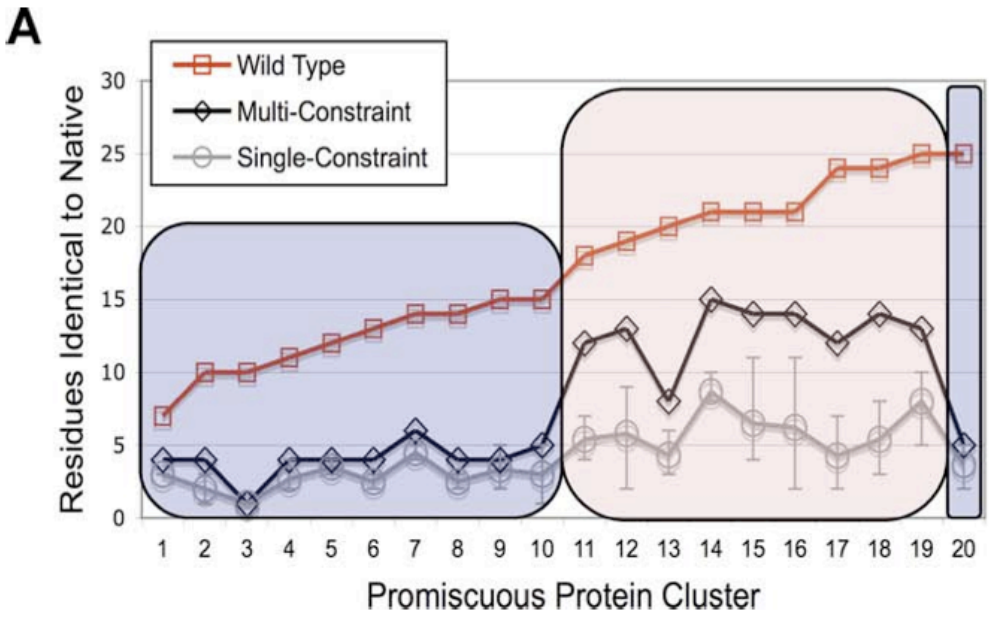
Figure 2.4 Single and Multi Constraint Structural Models for Two Ran Interface Sites
Computational models of interface regions around residues predicted to be optimal for binding each partner of Ran are shown (orange, 1A2K.pdb, yellow, 1I2M.pdb; green, 1IBR.pdb, purple, 1K5D.pdb; blue,

1WA5.pdb). Single-constraint predictions for residue 74 (A–E) (wild-type glycine) indicate compromise among the preferences of the five partners. Three partners (A,C,D), when optimized alone, prefer a residue with greater hydrogen bonding capabilities than the wild-type glycine. Steric constraints imposed by the remaining two partners (B,E) forced selection of the wild-type glycine by the multi-constraint protocol. Multi- constraint predictions for residue 76 are shown in panels F–J. The wild- type arginine is also chosen in all single-constraint predictions where it mediates an inter-chain hydrogen-bonding network (F,G,H,J). Single- constraint selection of leucine at position 76 for 1K5D.pdb is not shown.

In contrast to the compromised scenario described above, the bottom row of Figure 2.4 (panels F-J, structural multi-constraint models shown for Figure 2.3C, 2nd box, residue 76) depict a Ran interface residue that our simulations predict to be highly shared among all partners. Here the wild type residue, arginine, is correctly recovered by every single constraint simulation where it mediates an inter-chain hydrogen bonding network. This is the case for all partners except one (Figure 2.4I). Here the inter-chain interactions are formed largely by the aliphatic part of the arginine side chain, and design simulations favor a leucine residue. Hence for this interface position, where the multiple-constraint simulation also correctly selects the wild type arginine, there is little indication that recovery of this native amino acid is the result of compromises among the interaction partners. Interestingly, the Ran interaction partners depicted in Figure 2.4F and Figure 2.4G form very similar hydrogen bonding interactions with the wild-type arginine, although the partner proteins comprise different fold classes. This behavior of physio-chemically similar interactions formed by structurally distinct interfaces has been observed previously [48, 49].

2.6. Sequences selected by multi-constraint simulations can be substantially more native-like that single constraint sequences

We next investigated whether the trend of optimization for promiscuity using the multi-faceted scenario we observed for Ran was common in our dataset. In total, 65 separate single constraint optimizations and 20 multi-constraint optimizations were performed (Figure 2.2). Figure 2.5A shows that, over our entire dataset, sequences predicted as optimal by the multi-constraint protocol are more native-like than the sequences selected in the corresponding single-constraint runs (compare distance from red squares of black diamonds versus grey circles). There was only one instance (elastase in complex with inhibitors, promiscuous protein set #9) where the single-constraint optimization for binding one of the partners out-performed the multi-constraint protocol in native amino acid recovery.



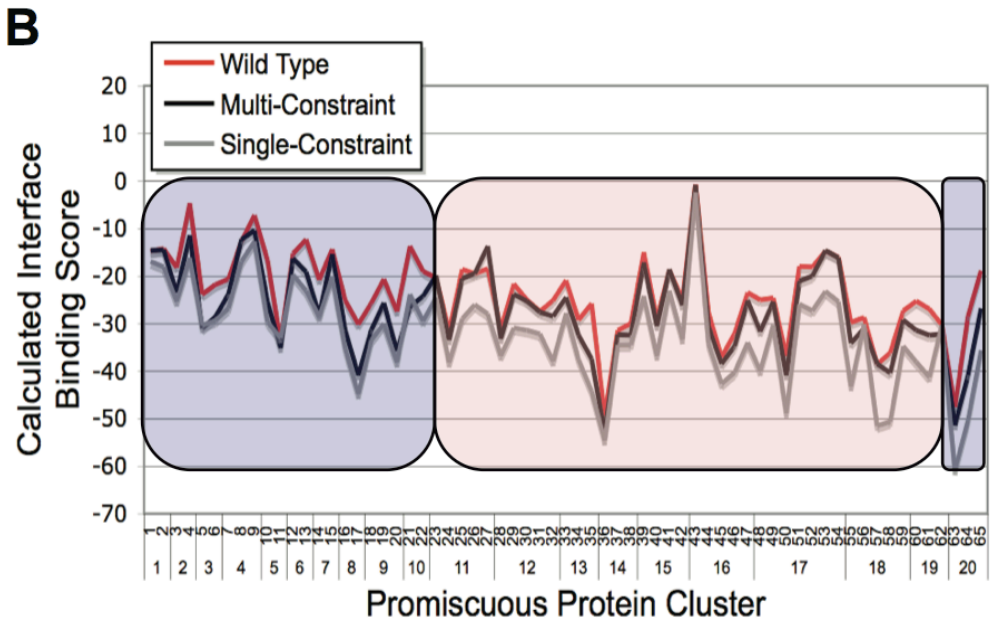


Figure 2.5 Comparison of Single- and Multi-State Sequence Recovery and Binding Scores
(A) The number of residues recovered as identical to native are plotted for each promiscuous protein (see Figure 2). For reference, the size of the shared interface is shown for each protein in red. For roughly half the dataset, (group II, pink shading), sequence recovery from the multi-constraint simulations (black) significantly out-performed the average single-constraint recovery (grey). The remaining proteins (group I, blue shading) showed similar native recovery regardless of whether sequences were optimized with respect to one or all characterized partners. Error bars represent the best and worst native sequence recovery in a single-constraint optimization. (B) Calculated binding scores of native (red), single-constraint (grey), and multi-constraint (black) sequences for each of the 65 complexes examined in this study (see Figure 2.2). Sequences selected by single- and multi-constraint optimizations often show a favorable decrease in binding score relative to native sequences for group I proteins (blue shading), while multi-constraint binding scores were close to native for group II proteins (pink shading).

Upon closer look at the pattern of interface residues recovered as native in each case, there seem to be two broad groups of multi-specific interfaces represented in the dataset. About half of the proteins comprised group I (blue shading, Figure 2.5A), for which the improvement in native sequence recovery in multi-constraint optimizations over single-constraint optimizations was small and total native amino acid recovery was low, regardless of interface size. As described for the shared scenario above, the low native sequence recovery could be due to all interaction partners binding via a few key residues, with the residues peripheral to these free to vary in sequence. This behavior is likely for several group I proteins including elastase, ovomucoid inhibitor and the SH3

domain complexes. These proteins bind their targets within a narrow groove or cavity and in addition a considerable fraction of interactions may be mediated through backbone contacts [50]. Low native sequence recovery in group I could also be influenced by inclusion of cross-species interactions (enzyme-inhibitor complexes and interleukin 6 receptor binding to mammalian and viral interleukin) as well as lack of sufficient constraints to fully specify the wild type sequence (see discussion below).

In contrast, for the other half of the proteins in our dataset (group II), sequence optimization over all characterized binding partners resulted in significant improvements in native sequence recovery compared with optimizations for binding to a single partner (pink shading, Figure 2.5A). Here, as described for the multi-faceted scenario above, the multi-constraint optimization procedure was able to "add up" differing amino acid preferences among partners. The resulting high recovery of native amino acids indicates that binding interfaces for proteins in this group are optimized for multi-specificity. Additionally, as compared with group I, group II proteins tended to use larger and flatter interfaces to mediate binding, were more likely to show high connectivity in protein-protein interactions networks, and bound interaction partners with a greater number of different fold types (Figure 2.2). Although generalizations of our conclusions are necessarily limited by the restricted size of our dataset of 20 proteins, a "multi-faceted" recognition pattern spread over a large interface may be a common strategy used by highly connected signaling hubs to bind diverse partners.

2.7. Binding scores of sequences selected by multiple constraint simulations are closer to native than those of single constraint sequences for group II interfaces

We have shown that for about half the multi-specific proteins in our dataset (group II) the multi-constraint designed sequences were substantially more native-like than single-constraint sequences (Figure 2.5A). According to our rationale outlined above, this suggested a significant level of optimization for multi-specificity in these interfaces. However, not all interface positions were predicted to be native-like, and native sequence recovery over the whole interface in multi-constraint simulations varied between 40 and 71% in this group.

Non-native amino acids could be chosen by our optimization protocol because they are predicted to be more favorable than the wild type residue or, alternatively, because a number of different amino acid types are allowed at a certain position without substantial energetic differences. To test whether the non-native interface residues selected by the design simulations were predicted to lead to significant interface stabilization, we compared the binding scores of sequences selected by the single- and multi-constraint protocols to the scores of the wild-type sequences. For both group I and group II, optimization for a single binding partner always resulted in a favorable decrease in predicted interface binding score (Figure 2.5B, grey line) relative to the wild type amino acid sequence (Figure 2.5B, red line). The binding score patterns for multi-constraint optimizations (Figure 2.5B, black line) however differed among the two groups: multi-constraint binding scores were often similar to single constraint scores for group I proteins (compare black and grey lines, blue shaded box) while for group II

proteins multi-constraint binding scores were much closer to those calculated for the wild type sequences (compare black and red lines, pink shaded box).

The division of our data set into two groups suggested by the native sequence recovery results (Figure 2.5A) were thus mirrored in the predicted binding score patterns for wild type and designed sequences (Figure 2.5B). Our simulations suggest that for group I proteins, where sequences and binding scores for single- and multi-constraint optimizations were similar, there might be non-native amino acids which could improve the promiscuous compromise and at the same time strengthen each interaction with each binding partner alone. In contrast, non-native amino acids selected for group II proteins in multi-constraint simulations are predicted to offer little improvement over the binding scores of the original wild type sequences; this confirms our notion of high levels of optimization for multi-specificity in this group. Interestingly, while our simulations sought solely to maximize binding affinity for each partner, and did not explicitly consider either the relative binding affinities among partners or that naturally occurring interfaces often need to be transient, incorporation of multiple constraints alone was often sufficient for our simulations to predict sequences with binding scores near or identical to that calculated for native sequences.

2.8. For all multi-specific interfaces, energetically important residues are generally optimized for binding

We next investigated, on a per-residue basis, at which interface positions our optimization protocols predicted native residues to be suboptimal. Experimental analysis of residues critical for maintaining binding with respect to a single interaction partner

have shown that often only a subset of the interface is comprised of key hotspot residues optimized for binding [40, 41] and that other non-hotspot positions may show a high degree of plasticity [43]. We thus wished to examine how often native residues were being recovered as optimal by our single and multi-constraint simulations at positions calculated to be energetically important "hotspots".

For each binding partner we calculated the per-residue score of the native residue at every interface position, and labeled sites with a native per-residue score of less than -2 as a predicted "hotspot". Next we calculated for each position the difference in score between the residue selected by each of our protocols and the score of the native residue (*Test for optimization*, Figure 2.1). We reasoned that small score differences (<1 score units; scores are parameterized to approximate kcal/mol [28]) should reflect that a given optimization protocol recovered the native (or energetically similar to native) residue during optimization and large score differences (>1 score units) should indicate the extent to which a non-native residue is predicted to improve binding affinity over native.

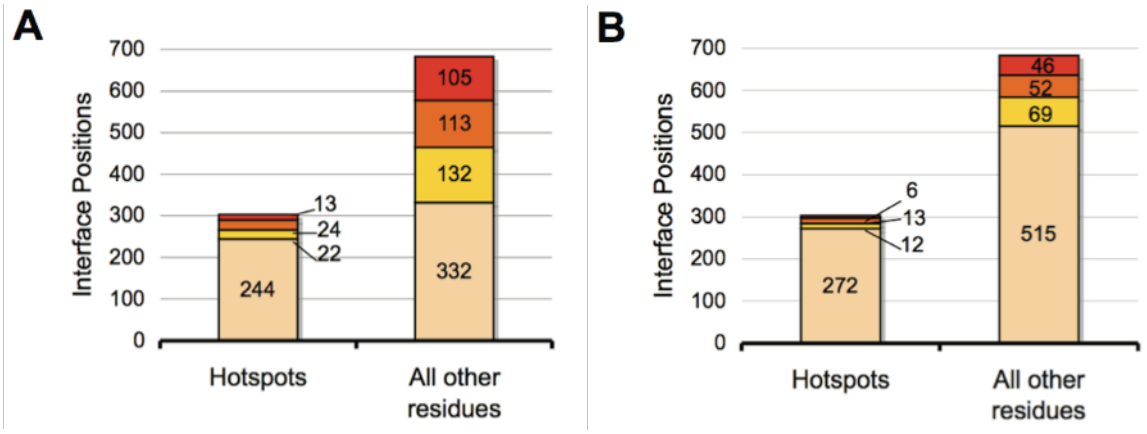


Figure 2.6 Distribution of Optimization in Promiscuous Interfaces
Predicted per-residue binding score improvements (relative to native) for sequences selected in single-constraint (A) and multi-constraint (B) simulations. Coloring indicates the magnitude of predicted improvement over native. Darker-colored bars (compromise value 1–1.5, orange; more than 1.5, red) indicate positions for which the simulation predicts a non-native residue to bind stronger than native.

Lighter-colored bars (compromise value 0–0.5, wheat; 0.5–1, yellow) indicate simulations recovered the native (or near-native) residue type. Whether optimization was in the context of single or multiple partners, positions calculated to be hotspots (see Section 2.4.4) consistently returned the native amino acid as optimal (244/303 and 272/303, for single- and multi-constraint simulations, respectively). In contrast, roughly half of non-hotspot interface positions were predicted as suboptimal for binding when each partner was considered separately (350/682), but only a quarter (167/682) were estimated to still be suboptimal in the context of binding multiple partners. Overall, the total number of interface sites for which improvements in binding scores could be found was significantly less for multi-constraint optimizations. Scores for the same residue position with differing binding partners are included in all plots.

At hotspot positions, whether optimizations were performed with respect to single or multiple partners, native (or energetically equivalent) residues were recovered for each partner with high fidelity (Figure 2.6A and Figure 2.6B, wheat bars, 244/303 and 272/303 for single- and multi-constraint optimizations, respectively). This inability to predict non-native residues scoring better than native at hotspot positions was seen for proteins in both group I and group II (Supplementary Figure A.1). In contrast, at non-hotspot positions, the native residue was predicted to be suboptimal (yellow, orange, red bars) with respect to binding a single partner in approximately half of all instances (Figure 2.6A, “all other residues”, 350/682). This is in agreement with experimental phage display data showing the native residue to often be suboptimal for binding at non-hotspot positions [43]. When considered in the context of binding multiple partners however, these same non-hotspot sites often are now predicted to be suboptimal in only 14% of all instances (Figure 2.6B, “all other residues”, yellow, orange, red bars 167/682). Thus we find that the need to maintain multi-specificity is imposing constraints primarily on non-hotspot residues, resulting in native residues being recovered more often at such sites as they become the “optimal compromise” for binding of multiple partners. This trend for increased recovery of native residues at non-hotspot positions during multi-constraint

simulations was much stronger for proteins in group II than in group I (Supplementary Figure A.1).

2.9. Distributions of shared and compromised interactions in promiscuous interfaces

Finally, we wished to estimate the extent of compromise each multi-specific protein in our dataset made in order to maintain binding to all its partners compared to the "ideal" interaction it could have if one a single partner was considered (see Section 2.2 and Figure 2.1A, *Estimate of Cost*). For each site within an interface, each partner was assigned a "compromise value" (ranging from 0-2). Compromise values were defined as the per-residue difference in score of the amino acid selected when each partner was optimized alone (single-constraint) and the residue selected at the same site when all partners were included in the optimization protocol (multi-constraint). The interface site itself was then assigned the largest compromise value seen among all binding partners.

For each position in the interface, this number should provide a rough estimate of the maximal amount of tradeoff paid by any partner due to the necessity of other partners binding via the same site (see Section 2.4.4 and Figure 2.1A). Small compromise values (0-1 score units) should indicate that all binding partners prefer the same (or similar) residue type as optimal, regardless of the presence or absence of other binding partners. Larger values (>1 score units) suggest that for at least one partner, a non-native amino acid is predicted to make more favorable interactions than the wild type, but may not be tolerated when preferences of all additional binding partners are considered.

Figure 2.7A shows, over our entire dataset, the percentage of sites within each protein interface calculated to have a compromise score between 0 and 0.5. These positions are predicted as essentially shared, in that no partner considered would have to give up potential gain so that other partners could fulfill their optimal interactions. While we observed a continuum ranging from interfaces calculated to have few completely shared interactions (all GTPases, Actin, Ubiquitin) to those for which the majority of interactions were shared (inhibitor complexes, SH3 domain), this analysis largely confirmed our earlier grouping of the multi-specific proteins within our dataset (Figure 2.7A, pink and blue symbols). A few group I proteins showed levels of compromise similar to that seen in group II. Interestingly, at least 2 of these proteins, Importin Beta (set #2) and CheY (set #4), were also calculated to be protein interaction "hubs" in our earlier analysis (Figure 2.2). These proteins may thus also employ a "multi-faceted" binding strategy, and the low native sequence recovery seen with the multi-constraint protocol is likely due to our computational prediction being under-determined (since we lack of structural information for a more complete set of binding partners). Likewise, we note that among the group II proteins, for IGG1-FC (set #15) many interactions were predicted as shared by all binding partners, a result that is consistent with an earlier structural analysis of these proteins by Delano et al. [48].

To illustrate the three-dimensional distribution of predicted compromises in multi-specific interfaces, we generated color-coded mappings of compromise scores. Representative maps for three promiscuous protein interfaces calculated to display high (Figure 2.7B, Ran), medium (Figure 2.7C, CheY), and low (Figure 2.7D, Ovomucoid Inhibitor) overall compromise are shown in Figure 2.7 (maps for the entire dataset are

given in Supplementary Figure A.2). Throughout our dataset, higher compromise scores often occurred along the periphery of a binding site, while highly shared residues tended to be more centrally located. While further analysis is needed, this could indicate strong, shared interactions with core hotspots may be necessary for each partner to bind, but that it is along the rim of the overlapping interface site where compromises among the binding partners have to be integrated in order to maintain multi-specificity. This is reminiscent of the idea that hotspot residues necessary for binding often occur in interface cores sequestered from solvent, whereas other non-hotspot parts of the interface, possibly around the rim, account for recognition [40].

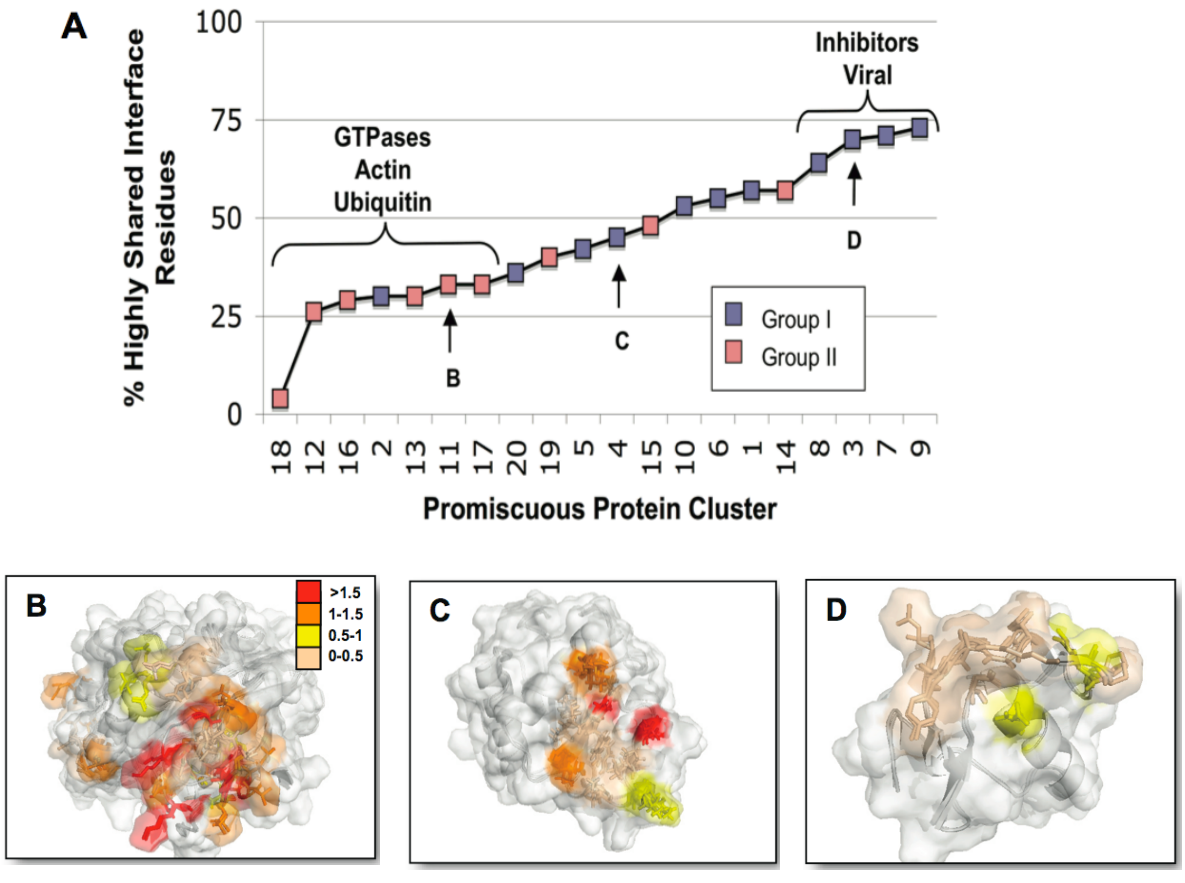


Figure 2.7 Distribution of Constraint Scores in Promiscuous Interfaces
Tradeoff at each interface position in our dataset was estimated by the per-residue difference in scores of amino acids chosen when each partner was optimized alone as compared with when all binding partners were considered in the optimization procedure. The percentage of interface sites displaying the lowest level

(0–0.5) of ‘‘tradeoff value’’ (see Section 2.4.4) is shown for all 20 proteins in our dataset (A). Such positions are predicted as highly shared, in that no partner considered had to ‘‘give up’’ potential gain so that other partners could fulfill their optimal interactions. Blue and pink shading denotes whether each protein was assigned to group I or II. Right-hand panels show color-coded mappings of constraint scores onto three promiscuous protein interfaces calculated to display high (B) (Ran set #11), medium (C) (CheY set #4), and low (D) (Ovomucoid Inhibitor set #3) compromise. Compromise values are colored as follows: 0–0.5, wheat; 0.5–1 yellow; 1–1.5 orange; >1.5 red.

2.10. Experimental verification of a non-native residue predicted optimal for multi-specific binding

Our energetic analysis suggests that many positions within naturally occurring multi-specific interfaces have been optimized for binding to multiple partners, while some native amino acids are predicted to be sub-optimal in the context of single or even multiple partners. Over the entire dataset, our multi-constraint protocol recovered the native interface residue as optimal for just under half (161/338) of all interface residues examined. Ultimately, experimental data are needed to verify whether choices of non-native amino acids by our multi-constraint optimization protocol are incorrect or underspecified predictions by our energy function, or whether the predicted choice would indeed strengthen binding for all partners.

In general experimental data validating binding affinities of sequences predicted by our single and multi-constraint simulations with all interaction partners were not available. However, we did observe one notable case where we could compare one of our predictions of an improved interface to direct experimental data. This occurred in the third domain of turkey ovomucoid inhibitor (set #3) at the key P1 position at which the inhibitor (or natural substrate) residue extends into a deep binding pocket. The predicted per-residue binding score at this site suggested that the wild-type residue was a hotspot crucial for maintaining binding with all partners, yet our multi-constraint protocol predicted a non-native amino acid residue to be significantly preferred over native by all

partners. As discussed above, prediction of a native "hotspot" residue to be suboptimal was an infrequent occurrence throughout our dataset (see Figure 2.6, "hotspots", yellow, red, orange bars).

Binding affinities for ovomucoid inhibitor mutants containing all 20 amino acids at the P1 position have been experimentally characterized for 6 different serine proteases [51]. This allowed us to compare the experimental preferences at the P1 position for the two serine protease complexes in our dataset (Chymotrypsin and SGPB, see Figure 2.2) with the computational predictions. The residue chosen at this site by the multi-constraint protocol, a phenylalanine, was ranked experimentally as the 3rd and 4th most favorable residue for Chymotrypsin and SGPB, respectively. There was no amino acid choice more favorable in common for both proteins and the native lysine residue was ranked 11th and 8th, respectively. We note that while our multi-constraint protocol correctly selected the optimal choice for binding the 2 characterized binding partners in our dataset, other amino acid types may be optimal for selectively binding different combinations of the 6 serine proteases studied. Interestingly, the P1 residue of ovomucoid inhibitor is known to vary significantly in nature, with 8 differing amino acid types occurring at this position in the 153 avian species analyzed [51].

2.11. Discussion

Our study uses a protein design method that can in principle be applied to computationally select amino acid sequences under any set of positive and negative constraints that can be defined by a fitness function. Here we have made comparisons between single- and multi-constraint predicted and naturally occurring sequences to quantify optimization and compromise in multi-specific interfaces.

Our analysis indicates that first, the protocol presented here is able to detect optimization for multi-specificity in promiscuous interfaces, as sequences and binding scores from multi-constraint simulations are closer to native than those obtained in single-constraint optimizations. Secondly, we identify two distinct mechanisms of achieving multi-specificity: (1) shared or low compromise interfaces where a small subset of interface residues have been optimized such that all binding partners utilize this set as hotspots and (2) multi-faceted or intermediate compromise interfaces where a far larger percentage of the interface has been optimized for multi-specific binding and each partner picks and chooses a subset of interface residue interaction with which to make key interactions.

Signaling proteins with large, flat interfaces fall clearly within the "multi-faceted" group II while enzymes, motif recognition domains, and receptors with smaller, narrower binding interfaces are often found within the shared group I. We speculate that the "multi-faceted" mode might have an evolutionary advantage for signaling interfaces, as here the chance that single mutations will deleteriously affect *all* binding interactions is reduced. On the other hand, a single mutation may substantially alter the *pattern* of interaction partners by now favoring certain interactions over others. In this way, multi-faceted interfaces may be more "evolvable" for new sets of interactions.

It is interesting to note that the ability to *a priori* predict binding sites from surface sequence conservation or surface cavity size has been shown to be easiest for proteins similar to those classified as "shared binding" by our methodology [47]. This is consistent with our observations, as in these cases there should be shared evolutionary pressure for conservation of key surface residues by all partners. In contrast, for proteins

predicted to display some degree of compromise among the differing binding preferences of their multiple partners, evolutionary pressures could differ depending on which subset of binding partners is most strongly selected for over time. Further, allowing each partner to pick and choose its own subset of interface amino acids for key interactions, as in the multi-faceted case, could necessitate large, easily accessible (i.e. flat) binding surfaces with a certain degree of conformational flexibility; this mechanism could hence partly explain why flat surfaces and conformational variability are frequently seen in multi-specific signaling proteins such as G-proteins [52].

We hypothesize that there should be significant differences in the ease with which binding specificities among partners could be rationally modified and/or small molecule inhibitors could be designed for proteins exhibiting the two modes of multi-specificity described here. The patterns of varying amino acid preferences among different binding partners revealed by comparing the single- and multi-constraint protocols suggest mutations at specific interface positions that could rationally change the specificity or promiscuity seen among binding partners. However, these same factors might make drug or small molecule design towards "multi-faceted" interfaces more difficult. For the group II interfaces in our set, the different partners display varying interface residue preferences (Figure 2.3C) and there may be a substantial number of constrained residues in each binding interface (Figure 2.7). Hence, proteins using this mode of interaction may have fairly distributed hot spots that are difficult to interfere with by a small molecule targeted to a single region.

A caveat of our study is that the generalizations may be somewhat limited because of the restricted size of our dataset of high-resolution structures. Further, the

results presented here are necessarily dependent on the quality of the scoring function used for optimizations. However, improvement in native recovery seen in multi-constraint simulations could not be directly due to energy function biases, as the same scoring function was used for all simulations. The ability of the Rosetta scoring function to predict energetically important residues has been analyzed previously [28]. We note that amino acid types for which our simulations consistently select the native residue as the best (optimal) choice for binding multiple partners include tryptophan, tyrosine, and arginine (Supplementary Figure A.3; the predicted amino acid frequencies for W, Y and R closely match the native distribution), amino acid types which have previously been shown to be energetically important in binding interfaces [40, 53, 54]. Interestingly, where allowed by steric constraints (for example at the interface periphery), we observed an increased selection in our simulations of larger amino acids such as tryptophan, arginine, and histidine, and against smaller amino acids such as alanine, threonine, and valine (Supplementary Figure A.3 and Supplementary Figure A.4). While this could be due to approximations in our scoring function, an alternative explanation could be that these non-native sequences would, at least in some cases, truly bind more strongly. An overrepresentation of large hydrophobic residues may have been selected against in nature to maintain protein solubility in the absence of binding partners. In addition, while our computational protocol optimizes binding score, naturally occurring transient interfaces may not necessarily have evolved for strong binding. The complexes between small GTPases and their exchange factors (GEFs) may be examples of interactions that need to be transient to fulfill their cellular function: in the case of the ARF1-Sec7 interaction, the fungal metabolite Brefeldin A inhibits signaling by stabilizing the

complex [55]. It may also be a general trend that multi-specificity must come at a cost of affinity [16]. Additional constraints not explicitly considered in our current protocol, such as selection at the level of on- or off-rates for complex formation could also account for differences in native and computationally selected sequences.

Lastly, we note that while the analysis presented here has focused on the ability of our simulations to identify the wild type amino acid, strict conservation of a single native amino acid over evolutionary time is rare, and the tolerance for substitution to differing amino acid types can vary between sites in an interface [47]. For example, for the multi-specific protein Ras we found two instances (Supplementary Figure A.4-12, positions 32Y and 67M) where we predicted the interface positions to be energetically important but failed to correctly recover the native amino acid. In both cases, the non-native amino acids selected by our multi-constraint simulations were among the evolutionarily tolerated set seen in a multiple sequence alignment (data not shown, generated as described in Section 2.4.6). A clear extension of our method is thus to not only predict optimal but a set of tolerated amino acid sequences for a given set of constraints (ELH and TK, unpublished data).

While we have applied the multi-constraint design protocol described in this work to examine whether and how promiscuous proteins are optimized for binding multiple partners, the methodology presented here is general and can be extended to analyze how any number of enumerable constraints (both positive and negative) affects sequence selection. A logical related analysis would be to characterize the sequence determinants of conformational flexibility where the input constraints would be stability for two or more different conformations. Further, the multi-constraint protocol introduced here is

not only predictive of naturally occurring amino acid sequences, but also allows for rational redesign of proteins with altered binding properties which could be instrumental towards understand the role of specificity in protein interaction networks as well as in the engineering of biosensors and new cellular pathways.

Chapter 3

Prediction of sequence diversity at a protein-protein interface using flexible backbone protein design

This chapter describes a method that uses flexible backbone move, inspired by coupled side chain-backbone motions observed in high-resolution protein crystal structures, to predict sets of amino acids tolerated at each interface position within the binding site of human Growth Hormone (hGH) with its receptor (hGHR). We compare our computational predictions to an experimental phage display dataset quantitatively mapping the sequence space of the hGH-hGHR interface and show that computationally predicted sequences to be enriched in functional members. Although the modeled structural changes are subtle, our results on predicting sequence plasticity suggest that backrub sampling may capture a sizable fraction of localized conformational changes that occur in natural proteins. The described method has implications for predicting sequence libraries to enable challenging protein engineering problems.

3.1. Introduction

Sequences of naturally occurring proteins show a remarkable robustness (neutrality) to mutations within a sizable "tolerated" sequence space, where many diverse sequence solutions are compatible with a given fold [56, 57]. This initial robustness [58] may help evolve new function [59], as sequence positions able to accommodate different amino acid residue types may be exploited to alter protein functionality. There is thus considerable interest in computational models describing the tolerated sequence space for a given protein fold and/or function, both to advance fundamental understanding of structure-function relationships [60, 61] and to engineer proteins with new properties [62].

Computational protein design methods which seek to identify low-energy sequences compatible with a target structure or interaction [5, 36, 60, 63], should in principle be able to capture at least some of the tolerated sequence variation of proteins. Accordingly, several methods have been developed to estimate the sequence space compatible with a given protein fold [25, 63, 64], but there are a number of theoretical challenges in evaluating the performance of any such method. Previous approaches have compared computational sequence predictions to multiple sequence alignments of protein families [56, 64-68]. However, evolved sequences have likely not sampled all tolerated amino acid combinations [43]. In addition, it is often difficult to determine whether given sequence positions have been under additional functional constraints not directly accounted for by design methodologies.

The ability to estimate tolerated sequence space also has practical implications for the engineering of proteins with new folds and functions. Despite several examples of

computational design successes, a re-occurring problem has been whether functional sequences can reliably be identified as top predictions. A promising approach to help overcome this problem and increase the success rate of design is therefore to combine computational predictions of the sequence space tolerated by a given fold and/or function with experimental library selection methods [69, 70]. This combined strategy can be used to significantly reduce the sequence space to be searched through experimentally. A key study showed that a library of sequences computationally designed to be more stable was also enriched in functional proteins [70]. This work suggests that computational methodologies can successfully sample within the sequence space available to folded proteins, but did not consider an explicit measure of the extent to which the full sequence space consistent with fold stability and function can be predicted computationally.

Here we aim to more directly assess the accuracy of computational methods for predicting the tolerated sequence space of folded and functional proteins. We make use of a recent study that screened phage display libraries to quantitatively map the sequence space of human growth hormone (hGH) able to bind human growth hormone receptor (hGHR) [43]. This dataset has several important properties: First, the diversity of sequences selected by phage display at this interface is considerable, making it a non-trivial test case for capturing tolerated sequence space. Second, the experimental pressures assayed, to fold and bind the receptor, can be directly mimicked computationally, without complications of other pressures acting on the sequences of evolved proteins. Third, the extensive phage display screening data can be compared to design predictions to assess the extent to which a hypothetical computational library is enriched in functional members.

It has previously been shown that protein design methods that neglect conformational adjustments, especially in cases where the input backbone design template is kept rigid, may restrict modeled sequence diversity. In contrast, models allowing for conformational diversity may broaden the sequence diversity accessible by protein design [64, 71, 72]. Therefore, we chose to compare computational predictions made using standard fixed-backbone simulations to those made using a model of conformational flexibility (termed "backrub") inspired by coupled side chain-backbone motions observed in high-resolution crystal structures [32, 33].

We find good qualitative agreement between the tolerated sequence space observed experimentally at the hGH-hGHR interface [43] and our predictions and show improvements of the flexible over the fixed backbone protocol. Based on these results, we suggest a flexible backbone computational protocol to design protein libraries enriched in functional members. Such a computational strategy of incorporating conformational flexibility and library design may broaden the use of computational methods for difficult engineering tasks such as constructing proteins with new functions.

3.2. Computational strategy for estimating the tolerated sequence space at protein-protein interfaces

We set out to develop a computational model for predicting tolerance to amino acid substitutions at protein-protein interfaces. Our overall computational strategy is illustrated in Figure 3.1 and outlined below. We reasoned that we needed, at a minimum: (1) a method to model local conformational changes (structural plasticity) that may result from sequence changes and binding interactions, (2) an efficient way to search sequence

space together with an easily computable folding and binding score for each sequence sampled, and (3) a measure to select the set of amino acid residues predicted to be tolerated at each interface position and compatible with the protein-protein interaction. We term these sets of amino acid residues a "tolerance profile" for each interface position. To test our computational strategy, we compare the computationally generated tolerance profiles to experimentally determined profiles [43] for each of the 35 positions in the ~1300A^2 hGH-hGHR site I interface (Figure 3.2A).

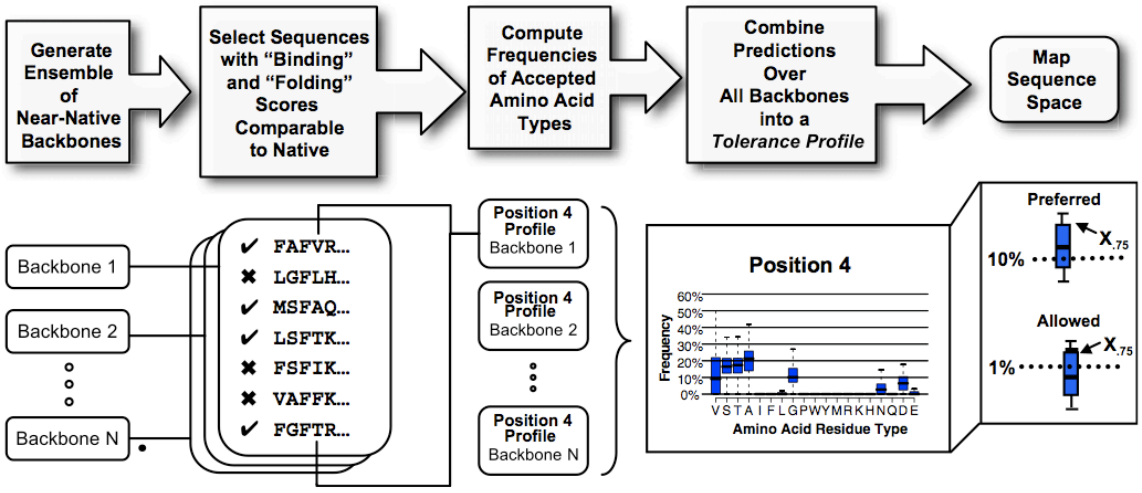


Figure 3.1 Schematic of the Computational Strategy for Predicting Interface Tolerance Profiles
Amino acid profiles were generated independently or each member of the ensemble by selecting for each backbone sequence predicted to have Rosetta 'binding'' (across-chain) and ''folding'' (within-chain) scores within a threshold score value of that of native hGH-hGHR interface sequence (checkmarks). Tolerance profiles over all members of the ensemble were combined and sets of allowed ($X_{.75}$>1%) and preferred ($X_{.75}$ >10%) amino acid residues were calculated for each interface position (see black dotted lines and arrows). Note that this protocol assumes independence of interface positions.

As a model for structural plasticity (requirement 1), we mimicked "backrub" motions inspired by coupled side chain-backbone motions observed in high-resolution crystal structures [32]. Briefly, each backrub move consists of selection of two residues (separated by 1 to 10 intervening residues) followed by a rigid rotation around the axis defined by the two C_α atoms of the selected residues. We generated an ensemble of 100 near-native backbones for both chains in the hGH-hGHR complex (Figure 3.2B; see

Section 3.3.3). As intuitively expected, the largest changes in C_α RMSD in the computationally generated hGH-hGHR ensemble occurred in loops connecting secondary elements. Average per-residue B-factors calculated for the computationally generated ensemble qualitatively reproduced experimental values for both hGH (Figure 3.2C) and hGHR (Figure 3.2D) (see also Supplementary Figure B.1).

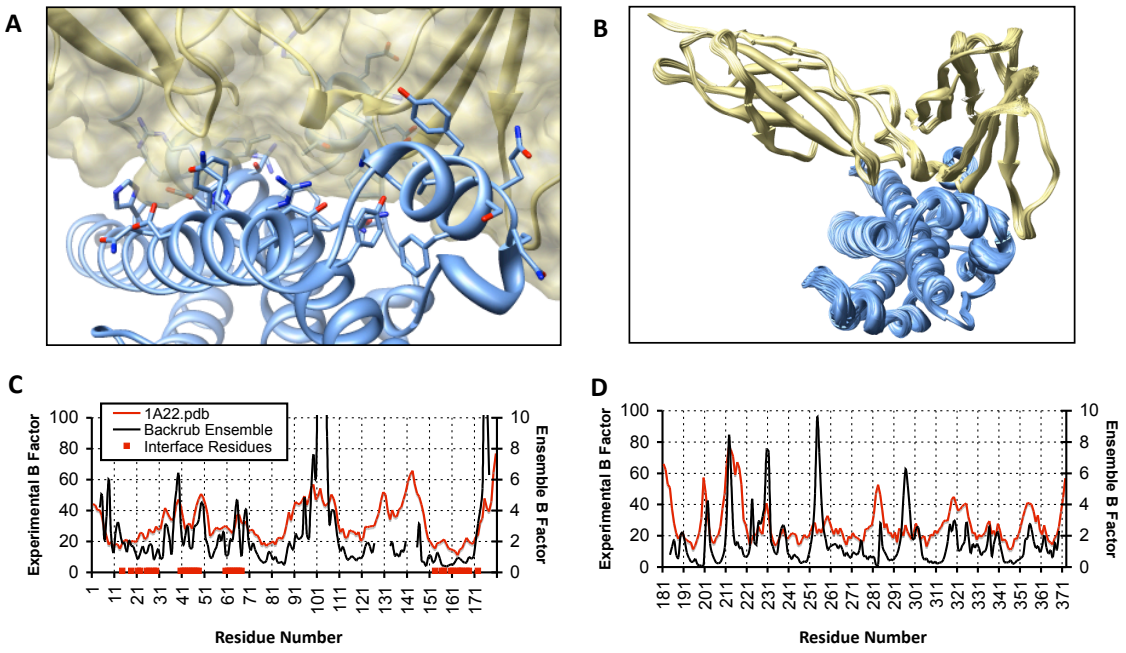


Figure 3.2 Overview of the hGH-hGHR interface and the Computationally Generated Backrub Ensemble
(A) Close-up of the 35 residues in the interface of human growth hormone (hGH, blue) with its receptor (hGHR, yellow). (B) Superposition of the 100 computationally generated backbones in the backbone ensemble. Comparison of hGH (C) and hGHR (D) B-factors in the crystallographic structure 1A22.pdb (red) and the hGH-hGHR backrub ensemble (black). Note that, due to our backrub procedure holding chain end-point residues rigid, ensemble B-factors have not been calculated for residues within five consecutive positions of a chain break or endpoint. Also note the differences in scales between the crystallographic and calculated ensemble B factors.

For sequence sampling and scoring (requirement 2) we chose to use the program Rosetta for protein design [68]. As a simplification, we performed design simulations on each backbone in the ensemble independently (Figure 3.1). Amino acid residue choices for the 35 hGH interface positions were sampled on ensemble members by performing several smaller independent simulations, each of which selected 5-6 hGH interface

positions to be targeted simultaneously for "design" and allowed to sample any of the 20 naturally occurring residues except cysteine (Section 3.3.5). This procedure directly mimicked the experimental phage display setup described in [43], and allowed our protocol to more efficiently search the large number of possible sequence combinations possible at the hGH-hGHR interface.

During each simulation, initially random interface sequences were scored by the Rosetta all-atom scoring function, which has been shown to predict the changes in stability upon point mutations in proteins and protein interfaces with reasonable accuracy [28]. Sequences with favorable binding (across hGH-hGHR interface) and folding (within the hGH fold) scores were then propagated using a previously published genetic algorithm (Section 3.3.5, [10] and [73]). Sequences sampled at any point during a simulation with both hGH-hGHR binding and hGH folding scores within a threshold value (a tunable parameter set to 1% of the score of the starting wild type sequence for all data shown) were saved and the frequency of appearance of each amino acid residue type in the selected sequences was recorded for every interface position (Figure 3.1). Profiles generated in this way for each backbone in the ensemble were then combined in box-plot format to visualize amino acid tolerances shared by many backbones ("flexible backbone" protocol) and compared with box-plots of 100 independent simulations on the fixed, crystallographic hGH-hGHR backbone ("fixed backbone" protocol). Note that the procedure described here assumes independence of the interface positions but can be extended to compute co-variation in designed sequences.

To address requirement 3, we defined measures to quantify which amino acid residues were predicted to be allowed at each interface site and which interface sites were

predicted to be most tolerant ("plastic") to amino acid substitution overall. We reasoned that subtle conformational changes in the hGH fold or at the hGH-hGHR interface might result in tolerance to certain amino acid types, and that the motions could vary depending on the interface position and the amino acid substitution being considered. Thus we hypothesized that some subsets of pre-generated ensemble members might be more or less appropriate for different interface positions and/or amino acid residue substitutions. To allow for this variability, we ordered the frequency with which each backbone member in the ensemble accepted a particular amino acid substitution at each interface position and selected the top 25% of all predictions for each site (this is represented by the $X_{.75}$ value; see box-plots in Figure 3.1 and Section 3.3.6 for details). Following this logic, we classified an amino acid residue type to be "allowed" by the *flexible* backbone protocol at a given interface position if it computationally appeared at a frequency of > 1% for each of the top 25% of ensemble backbone members "best suited" for the substitution considered ($X_{.75}$>1%; see Section 3.3.6 and Figure 3.1). Similarly, an amino acid residue type was considered "allowed" by the *fixed* backbone protocol when it appeared with a frequency > 1% on 100 independent design simulations on the crystallographic backbone. Amino acid types at each interface position for which the $X_{.75}$ value was ≥10% were considered to be "preferred" (see Section 3.3.6 and dotted lines and arrows in Figure 3.1).

Using these measures, we first show that both the flexible and fixed backbone protocols can successfully explore the tolerated sequence space at the model hGH-hGHR interface (Table 3.1 and Figure 3.3A). Second, we show that the flexible backbone protocol is better able to discriminate interface sites that experimentally showed near-

equal tolerance to substitution with many amino acid residues ("plastic" positions) from more "restricted" positions where fewer amino acid residues accounted for the majority of the sequences selected by phage display (Table 3.1). Third, for these restricted positions, we show that the flexible backbone protocol displays stronger qualitative (Figure 3.3B-C) and quantitative (Figure 3.4B-F; Table 3.1) agreement between computationally predicted and experimentally observed sequence tolerance profiles. Further, we provide structural examples and suggest possible explanations for how modeled backbone conformations might have resulted in improved predictions of tolerated sequences at several positions in the hGH-hGHR interface (Figure 3.5; Supplementary Figure B.2). Finally, we show that a hypothetical design library constructed by using our flexible backbone protocol would be enriched over a naïve library (consisting of residue types chemically similar to native, see Section 3.3.9) for generating folded hGH sequences that bind hGHR (Table 3.2/ Figure 3.6).

3.3. Methods

3.3.1. Source of experimental tolerance profiles.

Experimental data on comprehensive sequence mapping of the interface of human growth hormone (hGH) with its receptor (hGHR) were taken as published [43]. In order to be able to quantitatively screen sequence combinations of 20 amino acids, the 35 tested interface positions were experimentally screened in groups of 5-6 residues by using 6 separate display phage libraries each consisting of approximately 1010 unique members. The six residue groups, chosen in [43] to maximize the distance between individual interface positions as well as include no more than one known hot spot residue position per group, were as follows:

(1) 14M, 28Y, 47N, 61P, 171D, 179I

(2) 18H, 42Y, 62S, 65E, 164Y, 175T

(3) 21H, 29N, 45L, 60T, 67T, 178R

(4) 22Q, 43S, 66E, 167R, 176F, 183R

(5) 26D, 44F, 48P, 64R, 168K, 174E

(6) 25F, 41K, 46Q, 63N, 172K

3.3.2. Preparation of structures.

Starting atomic coordinates for all simulations were taken from PDB structure 1A22 (1:1 hGH-hGHR complex), which was stripped of all water molecules and prepared for design by an initial round of side chain minimization and optimization of hydrogens as previously described [73].

3.3.3. Generation of backbone ensembles.

All calculations were performed with the Rosetta full-atom scoring function, which is dominated by attractive and repulsive Lennard-Jones interactions, an orientation-dependent hydrogen bonding term [29], and an implicit solvation model [30]. In order to take into account sequence diversity which might require small backbone adjustments, an ensemble of "near-native" backbone conformations was generated from the minimized hGH-hGHR complex crystal structure by taking the lowest energy structures from 100 independent Monte Carlo runs using a new flexible backbone procedure termed "backrub" [32, 33]. A backrub move consists of rotating a peptide segment up to 40 degrees around an axis defined by the C_α atoms of two residues

separated by 1 to 10 intervening residues and subsequently optimizing the positions of the C_β and H_α atoms (branching off the pivot C_α atoms) according to the CHARMM bond angle potential as described in [33]. Backrub simulations were performed over all positions in both chains of the hGH-hGHR complex by repeatedly choosing random backrub beginning and end points and a rotation angle for a total of 10,000 moves, interleaved with side chain rotamer moves in the backrub regions. The resulting conformations of the 100-member ensemble had an RMSD to the native structure 1A22.pdb less than 0.4 Å.

3.3.4. Calculation of ensemble B-factors.

Average B-factors (the root mean square fluctuation, RMSF) for each C_α atom in the 100-member hGH-hGHR ensemble was calculated using Gromacs v3.3.2 and the function g_rmsf.

3.3.5. Generation of tolerance profiles.

Tolerance profiles for all 35 experimentally scanned positions of the hGH–hGHR interface were estimated from computational design simulations implemented in the program Rosetta. As described above, we performed designs on 6 independent groups (with a theoretical maximum of 106-107 sequence combinations per run), each of which allowed 5-6 interface positions ("designed positions") to vary among 19 amino acids (cysteine was excluded from all designs). During each design run, a genetic algorithm [73] was used to propagate an initially random population of 2000 sequences for five

generations for 100 independent simulations, scanning a total of 106 sequences. Runs with up to 30 generations gave essentially identical results (see Section B.1.1). This convergence is most likely explained by the fact that the experimentally screened groups were selected to contain positions spatially separated in the interface, reducing potential co-variation.

For each sequence selected by the genetic algorithm (including the native), Monte Carlo simulated annealing was used to minimize the Rosetta full energy function over the entire protein complex and find optimal rotameric conformations (chosen from the Dunbrack rotamer library [74] expanded around the chi1 and chi2 angles) of the selected sequence. All residues with a side-chain atom within 4Å of a designed residue were allowed to change their rotameric conformation ("repacked positions") while all other backbone and side-chain positions not considered for design or repacking were kept fixed. After optimized rotameric conformations were determined, each sequence was scored over the entire hGH-hGHR complex. The inter-chain score across the interface (all pair-wise scores i,j where atom i is found on hGH and atom j on hGHR) was used as an estimate for binding. The "binding" score was then subtracted from the complex score and the result was used as an estimate for "folding" (here the assumption is made that the score of hGHR for which the sequence is not changing is approximately constant). Sequence combinations with an inter-chain hGH-hGHR score within 1% of the "binding" score calculated for the native (starting) sequence and with an intra-chain hGH score within 1% of the "folding" score calculated for the native sequence were saved.

On average 150-5,000 sequences, dependent on the particular backbone used as well as the group of designed interface positions, passed both the binding and folding

thresholds and were used for inclusion in the computational tolerance profiles. Profiles from all "near-native" backbones were combined into a final computational prediction of tolerance at each interface position by plotting the median (as well as the 1st and 3rd quartile) frequency of occurrence of each amino acid residue (excluding cysteine) observed in each of the independent backbone runs. In order to estimate the variation in sequences selected in different runs on the same backbone, profiles for 100 independent simulations on the crystallographic backbone were also computed.

3.3.6. Selection of allowed and preferred amino acid types.

For all computational predictions, selection of allowed and preferred amino acid residue types were made using the third quartile ($X_{.75}$) value from the box-plots of tolerance profiles generated from either the 100-member ensemble or from 100 independent runs on the fixed, crystallographic backbone. At each interface position, an amino acid type was considered to be allowed and/or preferred if it occurred in the experimental phage display sequences (p_i) or computational predictions ($X_{.75}$) with a frequency of >1% or >10%, respectively.

3.3.7. Assignment of interface position tolerance levels.

In order to discriminate interface positions having one or more amino acid residue types marginally exceeding the 10% preferred threshold from other interface positions displaying moderate and/or strong selection for preferred amino acid residues, we devised the following metric. (Entropies were not used as direct comparison as the experimental assays and computational predictions allowed differing numbers of amino acids to be

sampled at each position; note cysteine was not computationally allowed). At every interface position, the amount by which each preferred amino acid type exceeded the 10% threshold was summed. If this value was <25, the interface position was labeled as having a high overall tolerance. Interface positions having values between 25 and 50 were classified as having medium tolerance and positions with a value >50 were classified as having low amino acid substitution tolerance. This process was implemented both for the sets of preferred amino acid residue types taken from the experimental phage display data, as well as for the sets computationally predicted by the fixed and flexible backbone protocols.

3.3.8. Generation of a computationally designed library.

At each interface position computationally predicted to have medium or low tolerance to substitution by the flexible backbone protocol, all amino acid residues identified as preferred were selected. While amino acids with strong biases in tolerance profiles could be selected for inclusion in a computational library by visual inspection (or setting the preferred threshold to a value differing from 10%), this process served to automate the predictions.

3.3.9. Selection of a naive library.

For each interface position, a "naïve" library was created by including all similar amino acid residue types in groups as defined below:

(1) [D,E,N,Q] (2) [R,K,H] (3) [L,I,V,M] (4) [F,Y,W] (5) [P,A,G] (6) [S,T]

3.4. Use of near-native backbone ensembles improves the ability to predict the overall tolerated amino acid sequence space

We used the protocol summarized in Figure 3.1 to computationally predict tolerance profiles for each of the 35 hGH interface positions examined in [43]. We first evaluated whether our protocol was able to generate sufficient sequence diversity to reflect the range seen experimentally. As the simplest measure, we compared the number of amino acid residue types experimentally observed at each interface position to the number computationally predicted with a small but observable frequency ("allowed" residues, experimental frequency pi or computational measure X.75 > 1%; see Table 3.1). The percentage of the experimental phage display sequences represented by the set of "allowed" residues selected by each protocol is also given in Table 3.1 and illustrated in Figure 3.3A.

Both flexible and fixed backbone protocols predicted sets of allowed amino acid residues that effectively covered the experimental sequence space for most of the 35 interface positions (Figure 3.3A; blue: flexible protocol; yellow: fixed protocol). On average, the fixed and flexible backbone protocols predicted sets of allowed amino acids that represented 89% and 92%, respectively, of the total phage display sequence space at each position. Notably, this result was not due to simply predicting all amino acid residue types to be allowed (Table 3.1), nor was it due to an overabundance of the native amino acid residue type in the phage display sequences (grey bars, Figure 3.3A).

Site	Allowed Amino Acid Residues						Preferred Amino Acid Residues‡						Measured ΔΔG‡‡
	Number			% Occurrence in Phage Sequences			Number			% Occurrence in Phage Sequences			
	Phage	Flex	Fixed	Phage	Flex	Fixed	Phage	Flex	Fixed	Phage	Flex	Fixed	
67T	8	8	7	-	85%	55%	4	6	5	81%	85%	55%	
171D	8	6	13	-	61%	84%	5	3	2	79%	49%	18%	★
176F	8	15	15	-	92%	92%	3	5	3	74%	86%	69%	★★
61P	9	8	10	-	73%	75%	1	4	7	52%	71%	72%	★
178R	10	17	13	-	100%	97%	2	5	1	80%	68%	19%	★★
183R	10	18	18	-	100%	100%	4	4	2	79%	2%	0%	
167R	11	16	17	-	94%	94%	2	1	1	47%	5%	5%	
172K	11	12	11	-	97%	97%	4	7	4	70%	85%	48%	★★
25F	12	16	16	-	96%	95%	5	4	3	68%	50%	47%	
179I	12	17	18	-	99%	100%	3	5	3	66%	59%	51%	★
60T	12	14	13	-	75%	75%	2	8	3	47%	48%	50%	
175T	12	14	15	-	100%	100%	4	7	2	61%	37%	10%	★★
43S	13	18	18	-	99%	99%	3	-	-	42%	-	-	
46Q	13	15	14	-	71%	70%	4	6	4	44%	38%	30%	
21H	14	13	15	-	87%	90%	3	3	4	50%	14%	14%	
45L	14	16	16	-	84%	90%	2	5	6	46%	62%	65%	★
64R	14	16	17	-	98%	98%	4	4	4	55%	43%	43%	★★
29Q	14	18	18	-	100%	100%	4	5	-	47%	30%	-	
44F	14	16	15	-	91%	88%	3	3	4	53%	44%	54%	
47N	14	18	18	-	97%	97%	3	2	-	44%	24%	-	
66E	14	18	16	-	96%	78%	3	-	-	41%	-	-	
14M	15	18	18	-	100%	100%	2	2	1	37%	11%	3%	
18H	15	18	18	-	100%	100%	3	4	1	38%	39%	17%	
28Y	15	17	16	-	95%	87%	2	4	4	40%	38%	31%	
41K	15	16	16	-	82%	82%	2	5	4	35%	42%	37%	
62S	15	19	19	-	100%	100%	4	2	-	50%	19%	-	
63N	15	16	15	-	92%	82%	4	3	1	52%	18%	2%	
22Q	16	15	15	-	78%	78%	4	3	3	51%	16%	16%	
48P	16	17	10	-	86%	51%	3	4	4	45%	33%	10%	
65E	16	19	19	-	100%	100%	3	-	-	50%	-	-	
164Y	16	16	16	-	95%	95%	2	-	-	29%	-	-	
168K	17	16	15	-	84%	77%	-	5	2	-	28%	15%	
174E	17	18	16	-	100%	92%	3	2	4	34%	10%	30%	
26D	18	18	18	-	100%	100%	1	-	-	11%	-	-	
42Y	18	18	18	-	100%	100%	1	5	3	10%	33%	23%	
Allowed	10^{39}	10^{41}	10^{41}	-	**92%**	**89%**							
Preferred							10^{15}	10^{17}	10^{11}	**50%**	**40%**	**31%**	
Restricted							10^{5}	10^{9}	10^{4}	**66%**	**48%**	**29%**	

Table 3.1 Allowed and Preferred Amino Acid Sets at the 35 hGH-hGHR Interface Positions
For each of the 35 hGH-hGHR interface positions (column 1), the number of amino acids determined to be allowed (see **Section 3.3.6**) by using either the experimental phage display sequences, the flexible backbone computational protocol, or the fixed backbone computation protocol are shown in columns 2–4. The percentage of phage display sequences represented by these sets of allowed amino acids are given in columns 5–7. Corresponding information for sets of amino acid residues selected as preferred is also shown (columns 8–13). Percentages over 75% and 50% are in red and blue font, respectively. Pink shading indicates interface positions identified by each method as ‘‘restricted.’’ Interface positions previously published as being hot ($\Delta\Delta G_{ALA\text{-}WT} \geq 1.0$ kcal/mol) and warm ($0.4 < \Delta\Delta G_{ALA\text{-}WT} < 1$ kcal/mol) spots are also indicated by ★★ and ★, respectively (column 14). Combinatorial sizes shown in the last 3 rows represent the total number of unique sequences if all combinations of allowed or preferred amino acid residue types were considered at all interface positions in a single library.

Although both flexible and fixed backbone protocols predicted similar numbers of amino acid residues to be allowed at most interface positions, there were improvements in predicting tolerated sequence space by using backbone ensembles at several positions (Table 3.1; compare blue bars and yellow diamonds, Figure 3.3A). For most of these interface positions the flexible backbone protocol correctly selected one or more amino acid residue types that occurred in the phage display sequences but were not predicted as “allowed” by the fixed backbone protocol (several cases are discussed further below). In contrast, there was only a single instance, position 171D, where the fixed backbone protocol appeared to be substantially closer to the experimental data (Figure 3.3A). However, direct comparison of the tolerance profiles generated by the flexible and fixed protocols at this position (Figure 3.4F, blue and yellow box-plots, respectively; experimental frequencies in red) shows that the flexible backbone protocol in fact improves predictions for the two amino acid residues most frequently observed experimentally (Ser and Asp). In this instance, the apparent better performance of the fixed backbone protocol is achieved at the expense of an overall higher bias for amino acid residues not observed experimentally.

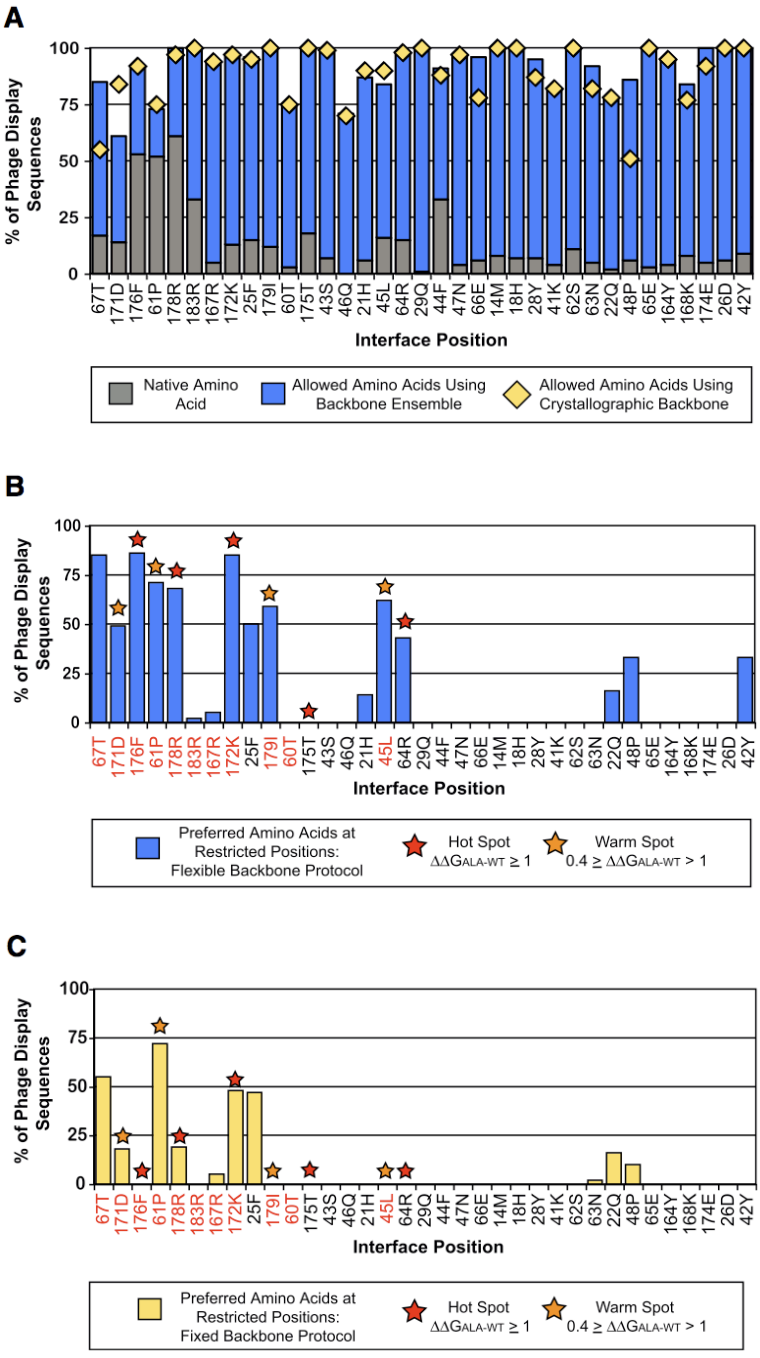


Figure 3.3 Ability of Fixed and Flexible Backbone Computational Protocols to Map the Allowed and Preferred Amino Acid Space

The percentages of phage display sequences accounted for by amino acids selected as allowed by the flexible (blue bars) and fixed (yellow diamonds) protocols are shown for all 35 hGH-hGHR interface positions (A). Interface positions are sorted by the number of amino acid residues observed experimentally (see Table 3.1). The percentage of sequences accounted for by the native amino acid is denoted by gray bars. The frequency percentage of phage display sequences accounted for by the preferred amino acid residues at positions selected as ‘‘restricted’’ by the flexible (blue bars) and fixed backbone protocol (yellow bars) are shown in (B) and (C), respectively. Positions indicated in red in (B) and (C) are classified

as restricted using the experimental phage display data. Hot spots and warm spots are denoted by stars, as indicated.

3.5. Design on a near-native conformational ensemble improves the ability to discriminate between restricted and plastic positions

We next sought to evaluate the ability of the fixed and flexible backbone protocols to identify small subsets of "preferred" amino acids residue types from the total set of amino acid residue types predicted to be allowed at each interface position. Table 3.1 shows, for each interface position, the number of preferred amino acid residue types observed experimentally and predicted computationally by both fixed and flexible backbone protocols (defined to be those occurring with a pi or $X_{.75}$ value >10%, respectively).

The sets of amino acid residue types selected as preferred by our computational protocols represented a significant reduction over the total allowed combinatorial sequence space (see estimates in last row, Table 3.1). On average, 3-4 "preferred" amino acid residue types were chosen at each interface site by the flexible and fixed backbone protocols, accounting for 40 and 31 %, respectively, of the experimentally phage display sequence space. We note, however, "preferred" amino acid types at some interface positions accounted for significantly more experimental phage display sequence space than others. We thus wanted to further test whether we could computationally classify such sequence positions based on whether they strongly preferred just a few amino acid substitutions.

To accomplish this, we assigned each interface position a "tolerance" value of HIGH, MEDIUM or LOW based on whether the sum total by which all its preferred residue choices exceeded the 10% frequency threshold was <25, 25-50, or >50 (see

Section 3.3.7). We call high tolerance interface sites “plastic”, as they are predicted to show an approximately equal experimental tolerance to substitution to a large number of amino acid residues (see for example Figure 3.4A). In contrast, we call positions labeled as having either a low or medium tolerance “restricted”, as typically only a small subset of preferred amino acid residues accounted for the majority of experimentally observed sequences. Sample tolerance profiles of interface positions labeled as plastic or restricted are given in Figure 3.4 (see Supplementary Figure B.2 for full dataset).

Over the entire dataset, the low/medium tolerance positions identified by the flexible backbone protocol favored interface sites experimentally determined as “restricted” (red coloring, Table 3.1 and Figure 3.3B) as well as positions previously determined by alanine scanning to be “hot” or “warm” spots [75, 76]. In contrast, the fixed backbone protocol misses 5 experimentally selected low/medium significance positions (176F, 183R, 179I, 60T, and 45L; see red coloring, Table 3.1 and Figure 3.3C) and incorrectly identifies more “hot” and/or “warm” spots as plastic/high tolerance. Further, the flexible backbone protocol selected restricted interface positions and sets of preferred amino acid types that accounted for, on average, a much larger percentage of the experimentally observed sequences than seen when using the fixed backbone protocol (29% and 48%, respectively for the fixed and flexible protocols).

The percentage of phage display sequences accounted for by the amino acid types selected as “preferred” by the flexible protocol showed improvements in coverage of the experimental sequences over those chosen by the fixed protocol at five interface positions: 67T, 171D, 178R, 172K, and 48P (compare blue and yellow bars in Figure 3.3B and Figure 3.3C). Despite the improved performance of the flexible backbone

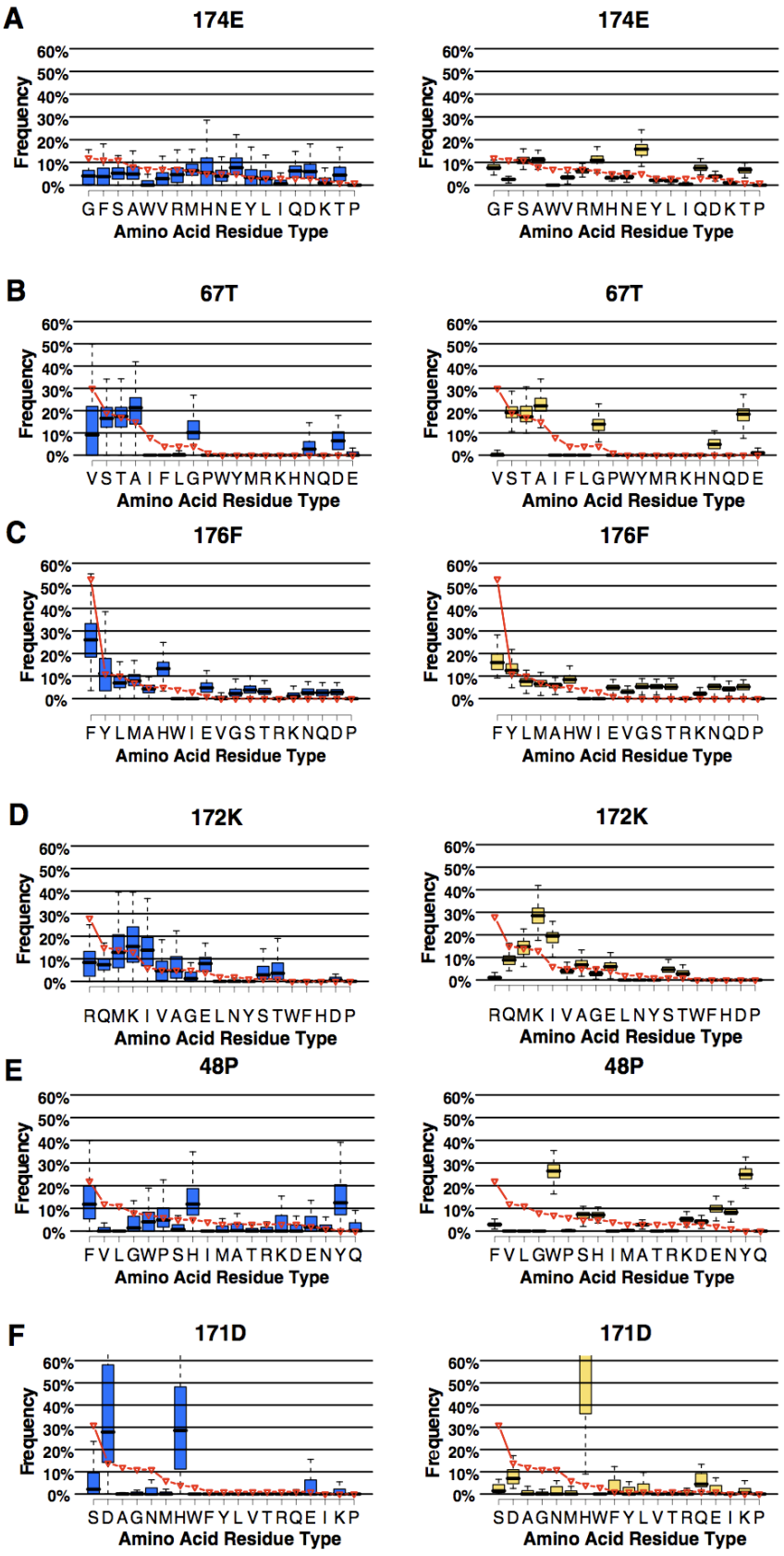
protocol at several interface positions, the tolerance profiles predicted by both protocols matched the experimental data poorly at four of the 35 positions (21H, 22Q, 167R and 183R; see Figure 3.3 and Table 3.1). These positions are discussed further in the following section.

3.6. Design on near-native ensembles improves the ability to qualitatively recapture experimental tolerance profiles

We showed above that, for many interface positions, the design protocols adequately model sequence diversity as well as preferentially identify interface positions with restricted amino acid preferences. Further we showed that both of these capabilities may be enhanced by design on near-native ensembles. To illustrate the origins of these observations, in this section we discuss in detail several representative examples of tolerance profiles generated by the two protocols at key interface positions.

Figure 3.4 compares representative tolerance profiles generated by the flexible (blue box-plots) and fixed backbone (yellow box-plots) protocols for seven interface positions (experimental profiles, red). Similar trends are seen throughout the dataset of 35 interface positions (Supplementary Figure B.2). While the overall shapes of the tolerance profiles were often qualitatively similar for the flexible and fixed backbone protocols, tolerance profiles generated using the 100 ensemble backbones generally showed wider variations in the frequencies of amino acid residues selected than seen for the 100 independent runs on the crystallographic backbone. This suggested that some subsets of backbone conformations favored certain amino acid residue substitutions not tolerated by other subsets of backbones. (There were 4 positions (see for example 171D and 178R,

Figure 3.4) for which larger variation, due to the presence of two distinct sub-distributions of selected sequences, was also observed for the fixed backbone protocol; see Section B.1.1 for details). The larger variation in the amino acid preferences observed in many of the tolerance profiles generated by the ensemble protocol directly resulted in the improved ability to distinguish plastic and restrictive positions and select sets of "preferred" amino acid residue types that more closely mimicked the experimental sequence preferences. For example, position 174E (Figure 3.4A) is correctly labeled as a "plastic" interface position by both the fixed and flexible backbone protocol. Note, however, that even while the two profiles appear qualitatively similar, the flexible backbone protocol shows some increased signal for at least two experimentally observed amino acid residue types (F and W). Likewise, the predicted profiles at positions 67T, 172K, 48P, and 171D clearly show experimental preferences for a few amino acid residues and are identified as "restricted" by both protocols (Figure 3.4B,D-F). In every instance, the flexible backbone protocol predicted the most frequently experimentally observed amino acid residue type to be preferred while the fixed backbone protocol did not (see amino acid types V, R, F, and S in Figure 3.4B, D-F, respectively). In contrast, similar amino acid residue types are selected as preferred by both protocols at position 176F (Figure 3.4C), but the overall profile shape appears somewhat flat when predictions are made on the crystallographic backbone (yellow box-plots). This results in this position being incorrectly labeled "plastic" by the fixed backbone protocol but being correctly identified as "restricted" when the flexible backbone procedure (blue box-plots) is used.


A
174E
Frequency
60%
50%
40%
30%
20%
10%
0%
G F S A W V R M H N E Y L I Q D K T P
Amino Acid Residue Type
174E
Frequency
60%
50%
40%
30%
20%
10%
0%
G F S A W V R M H N E Y L I Q D K T P
Amino Acid Residue Type
B
67T
Frequency
60%
50%
40%
30%
20%
10%
0%
V S T A I F L G P W Y M R K H N Q D E
Amino Acid Residue Type
67T
Frequency
60%
50%
40%
30%
20%
10%
0%
V S T A I F L G P W Y M R K H N Q D E
Amino Acid Residue Type
C
176F
Frequency
60%
50%
40%
30%
20%
10%
0%
F Y L M A H W I E V G S T R K N Q D P
Amino Acid Residue Type
176F
Frequency
60%
50%
40%
30%
20%
10%
0%
F Y L M A H W I E V G S T R K N Q D P
Amino Acid Residue Type
D
172K
Frequency
60%
50%
40%
30%
20%
10%
0%
R Q M K I V A G E L N Y S T W F H D P
Amino Acid Residue Type
172K
Frequency
60%
50%
40%
30%
20%
10%
0%
R Q M K I V A G E L N Y S T W F H D P
Amino Acid Residue Type
E
48P
Frequency
60%
50%
40%
30%
20%
10%
0%
F V L G W P S H I M A T R K D E N Y Q
Amino Acid Residue Type
48P
Frequency
60%
50%
40%
30%
20%
10%
0%
F V L G W P S H I M A T R K D E N Y Q
Amino Acid Residue Type
F
171D
Frequency
60%
50%
40%
30%
20%
10%
0%
S D A G N M H W F Y L V T R Q E I K P
Amino Acid Residue Type
171D
Frequency
60%
50%
40%
30%
20%
10%
0%
S D A G N M H W F Y L V T R Q E I K P
Amino Acid Residue Type

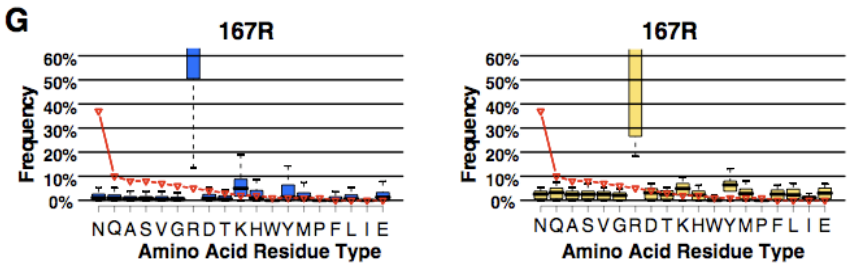


Figure 3.4 Comparison of Computationally Generated and Experimentally Determined Tolerance Profiles

(A–G) Computational predictions of amino acid tolerance profiles generated by either independent simulations on 100 near-native backbones (blue box-plots, left) or 100 independent simulations on the crystallographic backbone (yellow box-plots, right) for seven interface positions. Experimental tolerance profiles are denoted in each plot by red lines.

Lastly, we show a representative example of one of the four interface positions where both computational protocols poorly predicted the experimentally observed sequence preferences (Figure 3.4G; see also Table 3.1). At position 167, the native arginine was highly over-represented in the sequences chosen by both design protocols, leading to low signal at all other interface positions; this resulted in missing the experimental preference for asparagine. For position 183R, and 21H there was an incorrect computational bias for large aromatics (F, Y, H) that resulted in missing the strong experimental preferences (Supplementary Figure B.2F,O). Position 22Q (Supplementary Figure B.2AB), shown to be highly plastic and tolerant to amino acid substitutions experimentally, was incorrectly classified as restricted by both computational protocols. These prediction errors may be due to simplifications in our electrostatics model as well as general difficulties in modeling interactions of charged and polar residues in interfaces. We note that while our computational protocols failed to correctly predict highly preferred amino acid types at all four positions discussed above, the experimentally preferred amino acid residue types favored were nevertheless computationally predicted as “allowed” in each case.

3.7. Structural Analysis of Modeled Backbone Changes

As we saw some significant improvements with the flexible backbone protocol, we investigated the modeled structural consequences of backbone flexibility at three positions: 171, 67 and 48.

In the first case (Figure 3.5A), backbone flexibility around the D171 site appears to enable the formation of an altered polar interaction network around the D171 site. These changed side chain interactions may favor replacing the aspartate at position 171 with serine, which is the most frequently observed amino acid residue at this site by phage display and included as "preferred" amino acid by the flexible backbone protocol (Figure 3.4F). Consistent with our predictions, some of the modeled altered polar interactions are also observed in a crystal structure of a hGH variant with 15 point mutations in complex with 2 hGHR molecules which included the D171S hGH mutation (pdb code 1KF9; Figure 3.5B). We should note that the backrub flexible backbone method used here only models relatively subtle backbone conformational changes and may therefore not capture some of the larger structural variations such as those observed in the 1KF9 crystal structure [77].

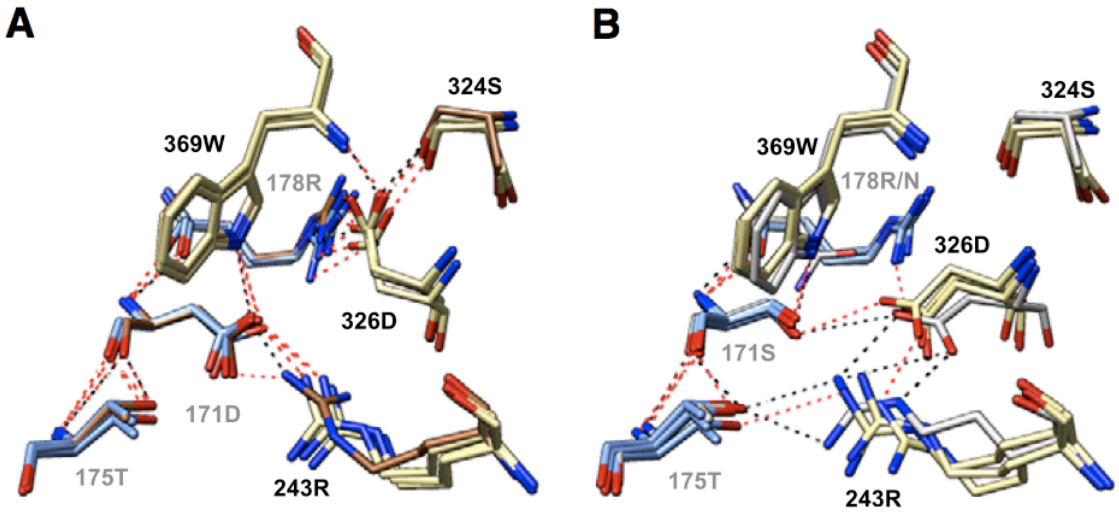


Figure 3.5 Structural Illustrations of the Possible Consequences of Backbone Flexibility
(A) Comparison of side-chain placement observed in the minimized crystallographic hGH-hGHR complex (PDB ID code 1A22, brown) to side-chain placement seen in three ensemble members (yellow, hGHR; blue, hGH) which prefer an aspartate at position 171. Residues with gray labels belong to hGH; residues with black labels belong to hGHR. Black and red dotted lines indicate hydrogen bonds in PDB ID code 1A22 and the ensemble members, respectively. (B) Ensemble members that prefer mutation of aspartate to serine at position 171 are compared to a crystallographic structure of an hGH-hGHR complex with 15 interface mutations, including 171S and 178N (PDB ID code 1KF9, gray). Hydrogen-bonding interactions (black and red dotted lines for PDB ID code 1KF9 and the ensemble members, respectively) and side-chain rearrangements observed for the ensemble members with the 171D-to-171S substitution mimic the interactions observed in the experimentally determined structure. Note the difference in the conformation in 326D. Modeled subtle backbone flexibility may enable this change in the 326D rotamer, which favors the D171S substitution

In a second example, the flexible backbone protocol correctly predicts that valine should be allowed at position 67 whereas the fixed backbone model misses this tolerated substitution. In this case, the hydrogen bonding interaction between 67T and a backbone nitrogen seen in the 1A22.pdb crystallographic backbone is frequently lost when modeling subtle (0.4Å C_α RMSD) backbone flexibility, favoring the experimentally observed threonine to valine substitution (Supplementary Figure B.3A).

In the third case, using the crystallographic structure 1A22, the substitutions 48Y and 48W are predicted to be stabilizing at the native 48P site. However, predictions made using the flexible backbone ensemble favor the observed 48F substitution over the larger tyrosine and tryptophan mutations. This could be due to side-chain rearrangements of 49Q and 274Q enabled by backbone flexibility, resulting in a somewhat smaller pocket surrounding positions 48 (Supplementary Figure B.3B).

As illustrated in these examples, our results predict modeled backbone flexibility can have a variety of consequences, including alternative side chain hydrogen-bonding networks as well as increasing and decreasing the volume at a given interface site. Similar results showing that backbone flexibility can lead to both prediction of increased or decreased side-chain flexibility have also been observed in an application of backrub ensembles to modeling side-chain flexibility and comparison to side-chain order

parameters determined by NMR [78]. It should be noted that in the case of hGH we cannot directly assess the accuracy of the flexible backbone predictions in Supplementary Figure B.3**,** as we lack crystal structures of point mutations. However, our analysis of the D171S mutation (Figure 3.5) appears consistent with structural data, as discussed above. In addition, we previously validated backrub-generated predictions of point mutant structures on a dataset of more than 2000 cases where experimental structures of both the wild-type and the variant were available [33].

3.8. Performance of computationally selected amino acid residues in a design library

Finally, we evaluated the ability of a hypothetical design library, consisting of the "preferred" amino acid residue types at positions selected as "restricted" by our flexible backbone protocol, to recover a large percentage of the experimentally observed phage display sequences by comparing it to two other hypothetical libraries. As a "lower" bound, we considered a naïve library, consisting of amino acid residues chemically similar to the native amino acid (Section 3.3.9). As an "upper" bound we considered a "perfect" design library, having the same number of amino acid residue types at each position as in the design library but consisting of the amino acid residue types occurring most frequently in the phage display sequences. The amino acids selected in each library as well as the sum of the frequencies of the selected amino acid residue types in the phage display sequences are given in Table 3.2.

Interface Position	Amino Acids Selected				% Occurrence in Phage Sequences			
	Native	"Perfect" Library	Computational Library	Naïve Library	Native	Native + "Perfect"	Native + Computational	Native + Naïve
67	T	V S A I G	A V S G D	S	17%	93%	85%	36%
171	D	S A	H S	N E Q	14%	57%	49%	27%
176	F	Y L M H	Y H M L	W Y	53%	86%	86%	68%
61	P	V S A	A G S	A G	52%	76%	71%	63%
178	R	H Q K T	M K Q L	K H	61%	89%	68%	83%
183	R	A G K T	F Y H M	K H	33%	85%	35%‡	43%
167	R	--	--	K H	5%	5%	5%	9%
172	K	R Q M I V A	M I R Q A V	R H	13%	86%	85%	41%
25	F	Y W A	Y W R	Y W	15%	58%	50%	47%
179	I	V Q T M	H V K M	V L M	12%	79%	59%	52%
21	H	G S	F Y	R K	6%	45%	14%	7%
45	L	F M W Y	Y M F H	I V M	16%	68%	62%	32%
64	R	F Y L	F K H	K H	15%	55%	43%	25%
22	Q	V E D	H N S	N E D	2%	43%	18%‡	33%
48	P	F V L	F Y H	G A	6%	51%	33%	17%
42	Y	V F K D	F H K R	F W	9%	42%	33%	21%
Combinatorial Library Size		**6×10^9**	**6×10^9**	**9×10^7**				
Average % Occurrence					**21%**	**64%**	**50%**	**38%**

Table 3.2 Comparison of Computational, Naive, and Perfect Design Libraries
Sixteen hGH-hGHR interface positions (column 1) and amino acid residue types (column 4) selected for inclusion in the computationally designed library (see **Section 3.3.8**) are shown. The native amino acid residue type for each selected position is given in column 2. For comparison, a perfect library (column 3) with the same number of amino acid residue types as chosen by the computational library and a naive library (column 5) are also shown. The experimentally observed frequency for the native residue (column 6) as well as the sum of experimental frequencies for all library amino acid residues (including the native) are shown for each library (columns 7–9). Amino acid residue types are shown for the design library in the order by which they were selected computationally (column 4), excluding the native amino acid residue type. For position 167, only the native R was selected as ''preferred'' by both the fixed and flexible protocols. Superscripts for positions 183 and 22 in column 8 indicate that the native amino acid residue type was not selected as preferred in the computational library. Average percentages of phage display sequences sampled over all 16 positions are indicated below, together with the size of each library. ‡Percentage differs from that given in Table 3.2 due to inclusion of the native amino acid residue type.

The library designed by the flexible backbone protocol performs about equal or better than a "naïve" library at 13 of the 16 selected interface position (Figure 3.6, compare blue bars with green triangles; grey bars indicate the frequency of occurrence of the native residue). The overall better performance of our design library suggests that the computational protocol is able to suggest non-intuitive amino acid substitutions. For instance, our predictions correctly identify the strong experimental preference for

phenylalanine at positions 45L, 64R, and 48P. Likewise, the frequent observation of serine at position 171D, as well as methionine and glutamine at position 172K, are predicted computationally (Table 3.2)

Comparison of the flexible backbone library to a "perfect" sequence set (Figure 3.6, red circles) shows that, for the size of the library, our algorithm is making near optimal selections of amino acids in the majority of cases. Over the 16 interface positions selected, the computationally designed library identified sets of amino acid residues coming, on average, within approximately 14% of the observed sequence space covered by a "perfect" library of the same size. This can be compared to the "naïve" library, which, on average, only comes to within 26% of the perfect library. In total, 38, 50 and 64% of the observed sequence space would be covered by the naïve, designed and "perfect" libraries, respectively (see Table 3.2). While the design library did occasionally select amino acid residue types never observed in the phage display sequences (predominately for the previously discussed positions 183R and 21H), we note that this may not be detrimental in actual experimental library selection applications, as long as other tolerated residue choices are included at the position in question.

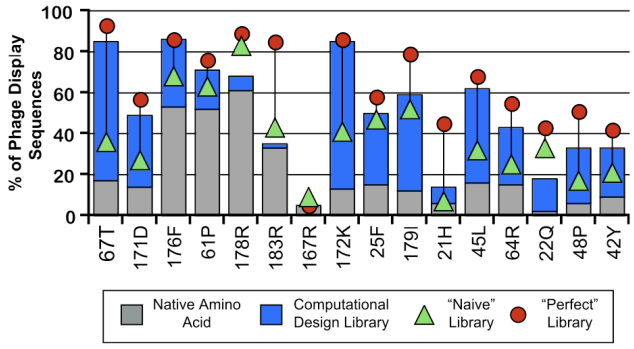


Figure 3.6 Performance of a Hypothetical Computationally Designed Library
The percentage of experimentally observed sequences accounted for by the amino acids selected in a hypothetical computational design library (blue bars, as in Figure 3.3B) is compared to a naive (green triangles) and perfect (red circles) library (see main text). For reference, the frequency of the native amino acid residue, which is assumed to be included in all libraries, is shown in gray in bar-plot form.

3.9. Discussion

We have developed a design strategy that selects sets of sequences consistent with an ensemble of backbone conformations and show that our protocol can (i) distinguish restrictive positions allowing only a few amino acid choices from highly plastic positions, (ii) predict tolerance profiles for restrictive positions that qualitatively match experimentally observed sequence preferences in many cases and (iii) be applied to design protein sequence libraries enriched in functional members.

There are two main new aspects of this work. First, we compare our predictions of protein interface sequence plasticity to experimentally determined sequence profiles obtained under multiple selective pressures we can mimic computationally. Second, we employ a new method to model structural plasticity in response to sequence changes and binding interactions inspired by coupled side chain-backbone "backrub" motions observed in high-resolution protein crystal structures [32, 33]. Although we cannot directly assess the structural accuracy of the near-native backbone ensembles, several lines of evidence suggest that backrub simulations are a useful model to represent local protein motions. Applying small backrub moves as implemented here has been shown to improve the prediction of side chain conformations in point mutant structures [33] and side chain order parameters obtained from NMR measurements [79] while larger-amplitude backrub-type moves have been shown useful for modeling protein loop motions [33] and obtaining backbone ensembles consistent with protein dynamics as measured by residual dipolar coupling (Greg Friedland & T.K., unpublished data). Finally, in contrast to other methodologies used to study protein conformational flexibility such as molecular dynamics, generation of backrub ensembles is

computationally very efficient and can be used in iterative backbone sampling and design simulations discussed below.

Our study only uses a single dataset for assessment [43], and it would be ideal to have additional protein interfaces with equally comprehensive experimental characterization. We would like to note, however, that the applied methodology does not involve substantial optimization of parameters to the test system, other than defining the threshold values mentioned above to guide the sequence tolerance profile analysis. The scoring function used here was derived using a large dataset of point mutations in both monomeric proteins [29] and protein-protein interfaces ([28]; this set did include the HGH-hGHR interface as one of 19 protein complexes) and applied successfully to recapture native-like sequences in a different dataset of 20 multi-specific protein complexes [73]. The backbone sampling method was parameterized independently [33] and not altered here. We also tested the robustness of our method to filtering of backbones and variation in thresholds (see Section B.1.1).

A simplification of the method in its described implementation is that it assumes independence of interface positions for compiling amino acid tolerance profiles. While we find good agreement with experimental data, this may be specific for this dataset, as the experimental study set out to avoid effects of co-variation as much as possible by choosing positions screened in each experimental library that were spatially separated in the interface. This is supported by the fact that we obtained similar results when designing one interface position at a time (while repacking a surrounding shell of 4Å) and computing tolerance profiles from Boltzmann-weighted scores over the backbone ensemble (data not shown). Independence of several positions in the hGH-hGHR

interface is further confirmed by a double-mutant-cycle analysis [80]. In such cases of minimal co-variation between positions, experimental selections targeting independent sites may be the most time-efficient approach if the main goal is to optimize interface affinity. Nevertheless, close agreement with extensive experimental datasets on tolerated sequence space for a protein interface, as shown here, is an important benchmark for assessing advances in computational design protocols.

Our method contains a number of other simplifications in the selection of preferred amino acids by compiling the amino acids from just the top 25% of the predictions at each site. We note here that we could not pre-specify subsets of backbones by examining either the starting score of the ensemble backbones with the native sequence or the final score of the backbones with the accepted modeled sequence, but instead needed to rely on the *difference* between these two scores. This occurred for two reasons. First, the ensemble backbones were generated with the native amino acid residues at every position in the interface. As we wished to provide opportunities for the ensemble to include backbones optimal for substitution with amino acid types considerably larger (or smaller) than native, some backbones with poor initial scores with respect to the native starting sequence were included in the ensemble. Secondly, the backrub procedure used to generate each member in the ensemble was global, allowing all residues within the hGH-hGHR complex to move, and likewise the ROSETTA scoring function was used to calculate a global score over all atoms in the hGH-hGHR complex. Thus an amino acid substitution could score better locally on a given ensemble backbone than on the starting crystallographic backbone, but the global score improvement from the substitution might not be sufficient to offset score changes due to

backbone movement distant from the interface position of interest. Both problems could be addressed by a protocol allowing local and independent optimization of the hGH-hGHR interface backbone following all appropriate amino acid substitutions. As this would be impractical for the number of sequences examined in this work, we instead relied on examining the frequencies with which each (pre-generated) backbone accepted each considered amino acid substitution.

While our current work suggests that simulations on backbone ensembles aid in capturing the experimentally observed sequence diversity in the hGH-hGHR interface, further developments may be needed in cases with substantial conformational coupling between backbone and side chain motions. Some important structural changes may only result from iteratively sampling backbone and side chain sequence moves rather than pre-generating backbone ensembles as mentioned above. For more general design tasks, such as estimating the sequence diversity of interacting residues in a tightly packed protein core, simulations may moreover be complicated by the presence of a rugged energy landscape with several local minima where co-variation cannot be ignored. Therefore, our methodology may need to be extended to directly compute correlated changes in the designed sequences [68]. Work on predicting structural effects of single point mutations supports the idea that such coupled changes can aid in the prediction of structural effects of sequence changes [33].

The ability to distinguish restrictive from highly plastic positions is important for several reasons: First, in a biological context, these predictions indicate which positions may be sensitive to mutations and which ones are not [81]. Second, a comparison of predicted plastic and yet evolutionarily invariant positions may provide testable

hypotheses for identifying amino acid residues that are conserved for reasons other than structural stability or binding affinity. These may include determinants of specificity where any amino acid residue would be allowed in the given interface but a specific residue may be necessary either for mediating binding with secondary partners [73] or selected for to minimize unwanted crosstalk with another undesired partner. Third, this information may be useful for engineering specificity determinants to form preferred and avoid undesired interactions within a set of possible partners [6, 10, 11, 81]. Finally, similar methods may also be useful to engineer enzyme active sites plastic enough to accommodate a variety of substrates; this substrate promiscuity may be an intrinsic property of naturally occurring enzymes and may confer evolvability [59].

Perhaps the broadest use of our described method is the design of sequence libraries enriched in functional members [69, 70]. This may be especially crucial for difficult engineering tasks where design methods are likely not accurate enough to predict a single functional sequence with high confidence. Here, combining computational design libraries with experimental screens may substantially increase the success rate over both testing individual sequences and unbiased screens [70]. In addition, the ability to select protein sequences for multiple criteria, such as fold stability combined with selectivity or multi-specificity towards several partners [73] could provide distinct advantages in applications where the goal is to generate designed proteins with several specified functional properties. While some desired properties, such as flexibility, interaction on- and off-rates, are difficult to include directly (and accurately) into design fitness function, designing a set of sequence and screening for desired properties may

yield a family of protein parts with a range of specifications desired for engineering and synthetic biology applications.

Chapter 4

Prediction of the mutational tolerance of HIV-1 Protease

Fast-evolving viral proteins such as HIV protease, under adaptive pressure to resist drug treatments provide ideal systems in which to examine mutational tolerance within the constraint of maintaining fold stability and native function. Here I test the ability of RosettaDesign to pro-actively predict the total sequence space tolerated by the HIV-protease fold in order to function under neutral (no selective pressure) conditions as well as mutations predicted to be sampled when a particular pressure (or drug) is applied. Other published work on HIV-Protease drug resistance typically examines a limited set of mutations near the binding pocket in the context of a single pressure (e.g. drug binding). In contrast, this work attempts to integrate the multiple the multiple structural and functional constraints (e.g. folding stability, dimer interface stability, native substrate binding and cleavage) acting on every sequence position of HIV-1 protease in order to make predictions of the total mutational space able to be tolerated. This work should provide useful insights into the mechanisms of mutational pathways, both for analyzing

viral evolution as well as predicting pathways to selectively engineer new protein functionality.

4.1. Introduction

Proteins appear to tolerate considerable sequence changes while still maintaining some of their native functional capabilities [82]. A fraction of the intrinsic sequence variability present in a protein population may become advantageous when conditions change to make a new secondary function more desirable [83, 84]. Directed evolution has been used to select new enzymatic capabilities without substantial loss of native activity [85-87]. Likewise, existing sequence variation in a population of functional proteins may provide the raw material from which resistance mutations can be rapidly selected after inhibitor treatment [88]. Thus, the ability to predict the extent to which proteins accommodate mutational changes could play a key role in forecasting the emergence of drug resistance or facilitating the engineering of new protein functions.

Here we test the idea of using computation to predict the accessible mutational space of proteins while preserving their natural function, as well as when selection pressures change. Using the fast-evolving viral protein HIV-1 protease as a model system, we develop a computational approach to predict mutations compatible with native fold and function but free to contribute to alternative functionality such as reduced inhibitor binding. To validate our model, we compare predictions of the sequence space sampled by HIV-1 protease under neutral and selective pressure to a collection of more than 50,000 HIV-1 sequences collected from untreated and treated HIV patients[89].

HIV-1 protease is a 99 residue dimeric aspartyl protease essential for correct processing and maturation of the HIV virus [90, 91]. In order for the HIV viral life cycle to be viable, the sequence of HIV protease must properly fold and dimerize, as well as recognize and cleave at least 10 endogenous pol and gag peptide sequences [92, 93]. Modeling the mutational space accessible to HIV protease is Due to the low fidelity and high error rate of HIV reverse transcriptase [94, 95], the HIV protease sequence mutates rapidly and the appearance of drug resistant mutation strains has greatly limited the overall effectiveness of HIV treatment [96]. It is known that higher levels of drug resistance are often correlated with an increasing number of protease mutations [97-99]. Several studies have investigated the effects of HIV-1 protease mutations located directly within the active site on inhibitor binding [100, 101]. While the contributions towards resistance of common non-active site mutations are generally less well understood, they have in a few cases been directly implicated in reduced inhibitor binding [102, 103]. However, a complete picture of the emergence of drug resistance may also require an understanding of the role active site and peripheral mutations play in overall viral fitness through exerting effects on the structure and function of HIV-1 protease.

We show integrating structural constraints on fold and dimer stability with functional constraints of peptide recognition can result in a computational prediction of a functionally tolerated mutational space very similar to that observed in HIV-1 clinical protease sequence variants [89]. The model we present successfully discriminates HIV-1 protease sites experimentally known to be intolerant to mutation with high fidelity, and narrows the possible mutational space down at sites more tolerant to mutation to a small, experimentally observed, subset of amino acid types. Surprisingly, without incorporating

any knowledge of the size or structure of possible inhibitors, the model predicts ~80% of known major and minor drug resistance mutations. Together, these simulations suggest that structural and functional constraints might be sufficient to predict functional HIV-1 protease sequence mutations from which specific drug resistance mutations may be selected during protease inhibitor treatment. This work suggests computational models of accessible mutational space, such as the one presented here, may prove to be generally applicable to modeling the evolvability of proteins by forecasting the emergence of mutations that can enable drug resistance or other new protein functionality.

4.2. Computational prediction of HIV-1 mutational frequencies based on structural and functional constraints

We set out to test whether modeling structural and functional constraints on sequence alone, without considering information of bound protease inhibitors, would enable prediction of the wide variety of escape mutations frequently observed after protease treatment. We hypothesized that in a population of fast-evolving viruses, such as HIV, there might already exist some significant sequence variability, even within a neutral setting, that might be selected from during inhibitor treatment. We used the program RosettaDesign [7, 28, 68] to estimate the effect of every possible point mutation on the stability of the HIV-1 protease fold, dimer interface, and recognition of endogenous substrate peptides. We varied the relative importance each mutation was required to make towards satisfying each of these structural or functional constraints in order to give two computational models. For predicting mutations likely to occur in the absence of any selective pressure, we chose a "neutral" model in which the importance of all constraints was weighted approximately equal (W_{FOLD}, W_{DIMER}, $W_{PEPTIDE}$ in

Supplementary Table C.3). In order to simulate pressure to accumulate mutations near the dimer and peptide binding sites, we also examined a "selective" model where dimer and peptide functional contributions were weakened relative to fold stability (by $^1/_4$ and $^1/_8$, respectively).

Proteins often adjust their conformations in response to sequence mutations, and backbone rearrangements have been shown to be important in several previous studies of HIV-1 protease point mutants [104, 105]. To account for such effects, we incorporated structural flexibility in two different ways. We first performed calculations independently on an ensemble of crystallographic backbone conformations. However, many of the crystallographic ensemble members originally contained one or more mutations as compared to the HIV-1 protease consensus "wild-type" sequence (subtype B, see Section 4.3.1). Thus we also later repeat our analysis for an ensemble of backbones computationally generated from structures crystallized in the absence of any mutation. Our computational protocol is illustrated for a single protease site in Figure 4.1 and described in greater detail below.

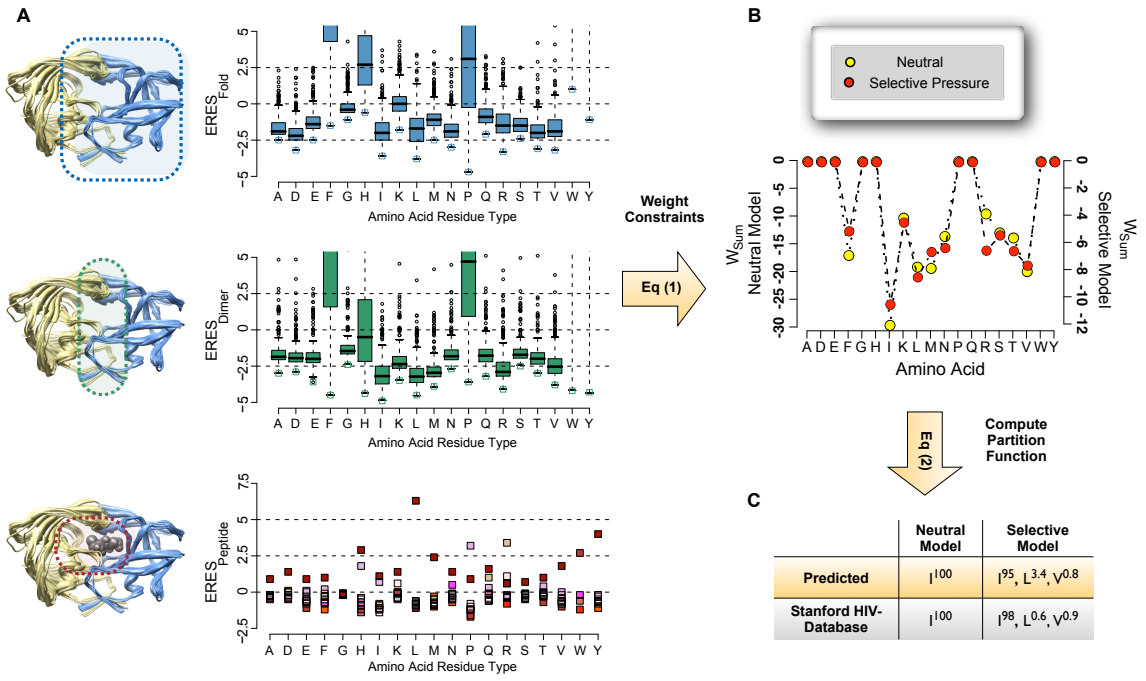


Figure 4.1 Mutational tolerance calculations for residue 50I.
19 amino acid types were computationally modeled onto a crystallographic ensemble of backbones with peptide absent (**A,** top, middle) or containing one of 10 endogenous substrate peptides (**A**, bottom). For each structure, the Rosetta++ per-residue contribution of each substituted amino acid types was recorded with respect to fold stability (blue, $ERES_{FOLD}$), dimer stability (green, $ERES_{DIMER}$) and peptide stability (red, $ERES_{PEPTIDE}$). The best ERES scores calculated among all members of the ensembles (**A**, open squares and filled squares; variation among ensemble members not shown for peptide calculations) was input into **Eq(1)** (see main text), using either "neutral" or "selective" model parameters. For both models, each amino acid was assigned a final weighted sum **(B)**. W_{SUM} values for all amino acids requiring more than one mutation at the nucleotide level from the consensus native amino acid type were set to zero. W_{SUM} values in **(B)** were then Boltzmann weighted to give predicted amino acid frequencies for each model **(C)**. The frequency of mutation observed within the HIV Stanford database HIV-1 protease sequences as site 50, before and after protease inhibitor treatment, is also given in **(C)**. Note that the scales of W_{SUM} value are of different magnitude for the "neutral" and "selective" models.

For predicting mutational effects on fold and dimer stability, we compiled a set of 263 dimeric HIV-1 crystallographic structures available in the protein database and removed all bound peptides and/or inhibitors from each structure (Supplementary Table C.1). For estimating the effect of mutations on interactions with protease substrates, we compiled a second set of 19 crystallographic structures and structural models of HIV-1 protease bound to 10 known peptide cleavage sequences (Supplementary Table C.2). The

sequences of all crystallographic structures were computationally reverted to the consensus wild type prior to making any calculations.

For each protein backbone, we then computationally mutated every position in the HIV-1 protease sequence, one position at a time, to each of 19 amino acid types (excluding cysteine) and optimized the side-chain conformations of neighboring residues using a rotamer library. The two sequence sites containing a native cysteine (67 and 95) and the site containing the catalytic aspartate (25) were omitted. For all 263 structures in the absence of peptide, we recorded the per-residue contribution of each mutation towards fold stability ($ERES_{FOLD,}$ all intra- and inter- residue contributions) and dimer stability ($ERES_{DIMER,}$ pair-wise score contribution with neighboring residues on the opposite dimeric chain). For each of the 19 structures containing a peptide, per residue score contributions each mutation made to the stability of each bound substrate ($ERES_{PEPTIDE,}$ pair-wise score contribution between the residue and all neighboring residues located on the substrate peptide) were also recorded. Each mutation was modeled simultaneously on both chains of HIV-1 protease, and ERES score contributions from both chains were taken into consideration by summing the values. In three peptide simulations, a portion of the HIV-1 protease sequence was itself a substrate (transframe region and HIV-1 protease cleave site (TF-PR), HIV-1 protease and reverse-transcriptase cleavage site (PR-RT), and the auto-proteolysis cleavage site (AutoP); see Supplementary Table C.2). For these simulations, each relevant mutation was modeled simultaneously onto both chains of the HIV-1 protease scaffold as well as onto the peptide backbone.

We observed that there was often a considerable variation in ERES score values calculated for amino acid substitutions modeled onto different backbones in the ensemble

of crystallographic structures. This was especially true for $ERES_{FOLD}$ and $ERES_{DIMER}$ scores (see, for example, Figure 4.1A, blue and green box-plots, respectively; variation in $ERES_{PEPTIDE}$ scores not shown). We thus chose to assign a single ERES value (denoted $\overline{ERES}$) for each constraint by taking the lowest (e.g. most favorable) predicted per-residue contribution for a given substitution observed over all ensemble members (see squares, Figure 4.1A). We reasoned that this process would allow the crystallographic backbone best suited for each individual substitution to be automatically selected.

After estimating the contribution of each amino acid substitution to fold stability, dimer stability and peptide binding, we integrated the functional and structural constraints predicted for each mutation by taking a weighted sum (W_{SUM}) of the calculated $\overline{ERES}$ values. In order to focus our analysis on mutations accessible by a single mutational step, we set the weighted sum term to zero for all mutations that involved more than one nucleotide change from the native (using MUTPROB, see Supplementary Table C.4). For all other amino acid types, the relative weighting of each constraint (W_{FOLD}, W_{DIMER}, and $W_{PEPTIDE}$) varied between our "neutral" and "selective" models as discussed above and detailed in Supplementary Table C.3**.** To set an overall mutational frequency, we favored the W_{SUM} value of the native residue type at each of the 99 HIV-1 protease sites by a constant amount ($FAVOR_{NATIVE}$). We also modeled a possible selection for protein solubility, which may oppose pure selection for protein stability, by disfavoring the W_{SUM} value mutations substituting a polar native residue with a hydrophobic amino acid type ($PENALTY_{POLAR \rightarrow HP}$). These calculations are summarized in Eq (1) and illustrated for a single residue site in Figure 4.1B. Parameter values are as given in Supplementary Table

C.3-**4** and details on the sensitivity of the overall results to these parameters are given in the Supplementary Figure C.1.

EQ (1)

$$\text{IF } MUT_{PROBi,j} = 0, \quad W_{SUM\,i,j} = 0$$

$$\text{ELSE}$$

$$W_{SUM\,i,j} = \frac{\overline{ERES}_{Fold_{i,j}}}{W_{Fold}} + \frac{\overline{ERES}_{Dimer_{i,j}}}{W_{Dimer}} + \frac{\sum_{k=1}^{10} \overline{ERES}_{Peptide_{i,j,k}}}{W_{Peptide}} + FAVOR_{Native} + PENALTY_{Polar \rightarrow HP}$$

where $MUT_{PROB_{i,j}}$ encodes whether amino acid type j can be reached by one nucleotide mutation at site i, W_{Fold}, W_{Dimer}, and $W_{Peptide}$ are the weights for each model constraint considered; $\overline{ERES}_{Fold_{i,j}}$, $\overline{ERES}_{Dimer_{i,j}}$, and $\overline{ERES}_{Peptide_{i,j,k}}$ are the best per-reisude constraint contributions at site i for amino acid j among all ensemble members, and $FAVOR_{Native}$ and $PENALTY_{Polar \rightarrow HP}$ are as described in the text.

In the final step of our calculations (Figure 4.1C), the weighted sum of ERES values was Boltzmann weighted at each HIV-1 protease site, giving rise to amino acid mutational frequencies predicted by both the "neutral" and "selective" computational models.

4.3. Methodology

4.3.1. Determination of mutational frequencies observed in patients

The "wild-type" HIV-1 protease sequence for subtype B, defined by the Stanford HIV database as the consensus sequence, is as follows:

PQITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMNLPGRWKPKMIGGI
GGFIKVRYDQILIEICGHKAIGTVLVGPTPVNIIGRNLLTQIGCTLNF

Frequencies of mutation observed in patients after protease treatment at each of the 99 HIV-1 protease sites were obtained online from the Stanford HIV drug resistance database (Genotype-Treatment Correlations/Treatment Profiles, see http://hivdb.stanford.edu/cgi-bin/PRMutSummary.cgi) by using the following settings: #

of PIs: 1-9, subtype: B, reference profile: subtype B untreated, exclude single occurrences: yes, include mixture: no, one mutation per person. Frequencies under a neutral setting were taken from the reference profile (subtype B, untreated) generated using the profile settings described above. Amino acid substitutions occurring in >=0.5% of sequences after protease inhibitor treatment are as given in Table 4.1.

4.3.2. Selection of crystallographic structures used for predictions of fold and dimer stability

386 crystallographic structures of HIV-1 protease were obtained from the protein databank (PDB) by using the search by sequence feature (Blast e-value 0.001) to retrieve structures with sequences similar to 1PRO.pdb (chain A). Structures which contained more than 12 mutations from the HIV-1 subtype B consensus sequence defined above, contained only one chain of the HIV-1 dimer, or with cysteine residues replaced (heteroatom residue codes ABA, CME, CSO, or DBU) were eliminated. Structures determined to be either HIV-2 protease, SIV, Rous sarcoma virus, or tethered dimeric HIV-1 were also eliminated. 263 dimeric HIV-1 crystal structures remained (see Supp. Table 1) and were collectively used as a crystallographic structural ensemble. The majority of structures contained 1-7 mutations (85% or 223/263) and had a crystallographic resolution within the range of 1.0 to 2.5 Angstroms (87% or 230/263).

4.3.3. Selection of crystallographic structures and generation of model structures used for predictions of substrate stability

16 dimeric, crystallographic structures crystallized with one of 7 endogenous peptide substrates were used for all calculations of substrate stability (Supplementary Table C.2). We generated structural models for each of the three peptide cleavage sequences without crystallographic structures (CTLNF-PISPI, PQITL-WQRPL, and VSFNF-PQITL) by computationally threading each peptide cleavage sequence onto each of the 16 crystallographic structures (sequence positions for which there was missing crystallographic density on any of the 16 peptide template structures were omitted), performing an initial round of side-chain minimization, and selecting the structural template for which the resulting Rosetta interface score (sum of score contributions over all pair wise interactions between residues i and j, where residue i was located on HIV-1 protease and residue j was located on the bound peptide) was lowest. 1MT9.pdb was found to be the best template for both peptides CTLNF-PISPI and PQITL-WQRPL, while 1F7A.pdb was selected as the optimal template for VSFNF-PQITL. Peptide interface scores for the resulting three models (-18.3 to -25) were within the range observed for 16 crystallographic peptides (-18.5 to -33).

4.3.4. Generation of backrub structural ensembles

Computational ensembles of "near-native" backbones were generated starting from one of 11 crystallographic structures originally crystallized with the native consensus sequences (1A8G.pdb, 1EBY.pdb, 1HXW.pdb, 1IZH.pdb, 1PRO.pdb, 1SBG.pdb, 1VIJ.pdb, 1VIK.pdb, 4PHV.pdb, 5HVP.pdb and 9HVP.pdb) by using a

previously described backrub protocol [33]. Briefly, the backrub protocol consisted of repeatedly selecting C_α atoms of two residues (separated by 1-10 intervening residues), performing a rigid body rotation of the selected protein segment (of up to 40 degrees), optimizing the location of related C_β and hydrogen atoms, and accepting or rejecting the backbone move based on the Rosetta scoring function and the Monte Carlo Metropolis criterion. Using the atomic coordinates of each crystallographic structure above as a starting conformation, 100 independent backrub simulations were run at two separate Monte Carlo temperatures (kT=0.6 and kT=1.2) until a total of 10,000 moves per simulation had been sampled. At each temperature, the lowest energy conformation sampled as well as the last conformation accepted during each simulation were saved and used to generate a computational ensemble of 400 backbone conformations per starting crystallographic structure.

Average RMSD values of backrub generated conformations relative to the starting crystal structures were dependent on the starting template and Monte Carlo temperature used, but typically and ranged, on average, from 0.2 to 0.6 for conformations generated at a kT of 0.6 and 0.3 to 0.8 for conformations generated at a kT of 1.2.

4.3.5. Energy Function and Preparation of Crystallographic Structures:

All computational calculations were performed using the Rosetta scoring function, which is dominated by attractive and repulsive Lennard-Jones interactions [29], an orientation-dependent hydrogen bonding term , and an implicit solvation model [30]. Simulations consisted of sampling and scoring side chains, taken from a rotamer library including the native amino acid PDB conformation and with additional rotamers around

the chi1 and chi2 side-chain torsion angles, [5] using a Monte-Carlo simulated annealing optimization protocol as described in [73]. Minimization and optimization protocols were with respect to the sum of the total score over all atoms within the dimeric HIV-protein (or dimeric HIV-protein substrate) complex.

In preparation for calculations of fold and dimer stability, all water molecules, heteroatoms, bound inhibitors or substrates and hydrogens present in the original 263 crystallographic PDB structures were removed, and hydrogen atoms were added as previously described [73]. An initial round of side-chain optimization was performed using the Rosetta scoring function as described above, keeping all amino acid identities and backbone coordinates fixed, while selecting for the optimal rotamer at each side-chain position from the rotamer set described above. After this initial minimization, all structures containing amino acid identities differing from the consensus HIV-1 subtype B sequence (see above) were computationally reverted to the consensus sequence and the structures were side-chain minimized a second time. In preparation for calculations of peptide binding stability, a protocol identical to that described above, except leaving all bound substrates present, was repeated separately for the 19 crystallographic and model structures listed in Supplementary Table C.2.

4.3.6. Selection of model parameters:

Optimal values for the six model parameters (W_{FOLD}, W_{DIMER}, $W_{PEPTIDE}$, $FAVOR_{NATIVE}$, $FAVOR_{POLAR\ NATIVE}$, and $PENALTY_{NATIVE\ POLAR,\ HYDRO}$) were selected by simultaneously varying each model parameter (Supplementary Table C.3) and computing predicted amino acid frequencies over 97 (non-cysteine) HIV protease sites. For each combination of parameters, the 99 HIV-1 residue sites were computationally

classified as displaying either low (1-5%), medium (5-20%), or high (>20%) mutation rates and the number of residue sites correctly matching the experimentally observed mutation rates was calculated. The percentage of sites correctly determined for each bin was then averaged and used to determine a parameter set for both the "neutral" and "selective pressure" computational models (see Supplementary Figure C.1 and Supplementary Table C.3).

4.3.7. Evaluation of ROC curves and AUC values:

True positive rates (TPR) and false positive rates (FPR) of mutation recovery were calculated by using the parameter values determined above for a "neutral" and "selective pressure" computational model (Supplementary Table C.3) and considering all mutations computationally predicted to occur at frequencies greater than or equal to the following cutoffs: 10%,9%,8%,7%,6%,5%,4%,3%,2%,1%,0.5%,0.1%, and 0.01%. Mutations determined to occur at an experimental frequency of >0% (e.g., greater than a single occurrence), >=1% and >=5% were considered as true positives in constructing ROC curves. ROC curves were constructed for mutations observed within the HIV-1 Stanford HIV-1 database after treatment with 1-9 protease inhibitors (Figure 4.3). AUC values were calculated for each ROC curve by implementing the trapezoid method.

TPR and FPR rates were calculated for two null models by considering at each of the 99 HIV-1 residue sites (1) the set of mutations one nucleotide mutation away from the native codon and (2) the set of all mutations to amino acid types chemically similar to the native, grouped as follows: (A,G,P),(D,E,N,Q),(F,W,Y), (L,I,V,M),(R,K,H),(S,T), see also Supplementary Table C.4**.**

4.4. Discrimination between low, medium, and high frequency mutational sites for "neutral" and "selective" models

We first examined the general ability of the neutral and selective computational models to discriminate HIV-1 protease sites tolerant to mutation from other sites observed to rarely or never mutate. We calculated how often each model predicted non-native amino acid types at every sequence site, and mapped these predictions to the HIV-1 protease structure **(**Figure 4.2B, neutral model; (Figure 4.2D, selective model; red, predicted mutational frequencies >20%, red; 5%-20%, orange; 1%-5% yellow). Structurally, the computational models qualitatively reproduced the pattern of mutational tolerance observed within the HIV Stanford Drug Resistance Database sequences by showing an increased frequency of mutation after protease treatment at the substrate interface and at the dimer flaps (Figure 4.2A, pre-protease treatment; (Figure 4.2C following inhibitor treatment).

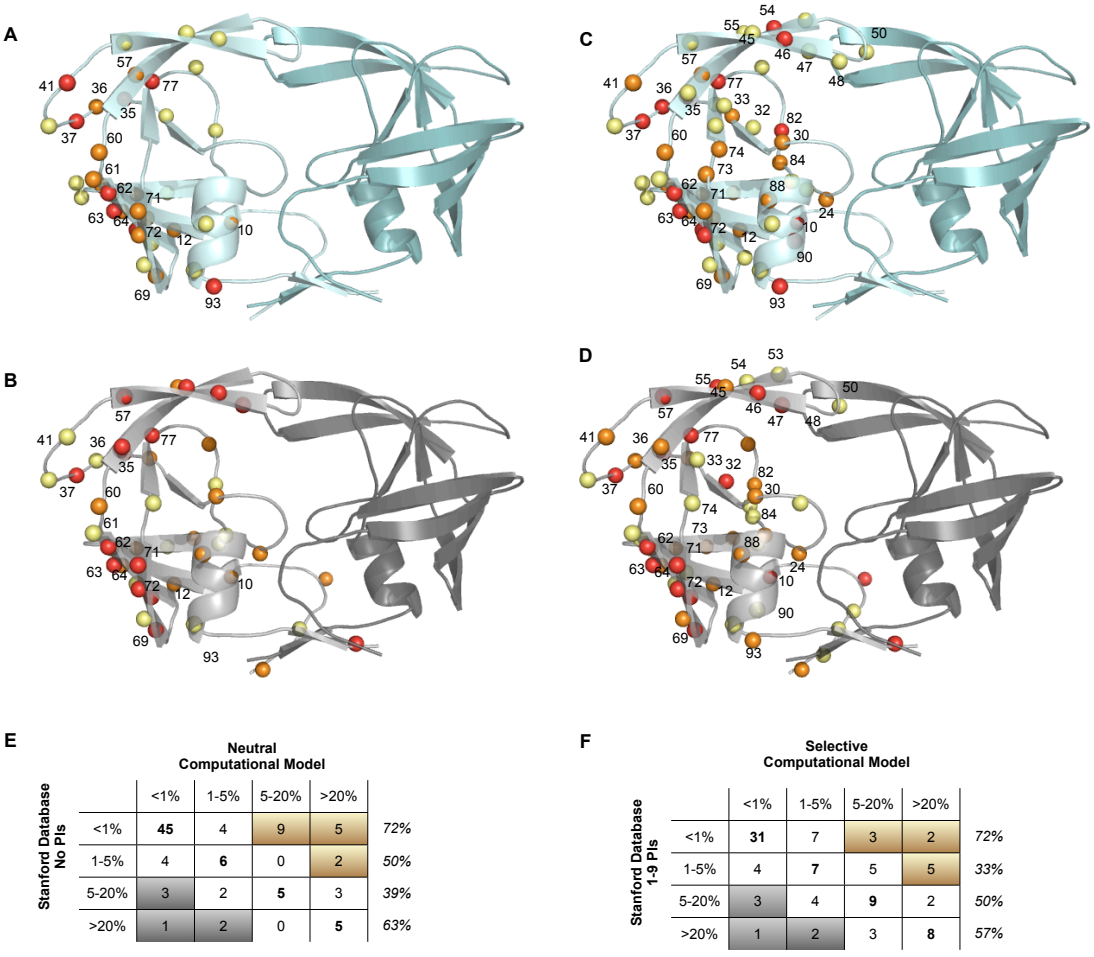


Figure 4.2 Predicted and Observed HIV-1 Protease Mutational Tolerances.
Total percentages of non-native amino acid types observed within the HIV-1 Stanford database sequences before **(A)** and after **(C)** protease treatment are mapped onto the HIV-1 protease structure at every site (excluding 25D, 67C, and 95C). Total percentages of non-native amino acids predicted by the neutral **(B)** and selective **(D)** computational model are also shown. Colored spheres indicate observed and predicted frequencies of mutation: >20%, red; 5-20%, orange; and 1-5% yellow. Sites with predicted or observed mutation to non-native amino acid types <1% are shown only in cartoon. For clarity spheres are shown only on chain A. Tables show the overlap between predicted and observed mutational frequencies. Grey shading denotes the number of sites for which the computational model under-predicted mutational tolerance with respect to the database sequences. Brown shading depicts sites predicted to have a higher tolerance to mutation than observed in the database sequences.

Both models correctly identified ~72% of the sites at which mutation in the database sequences was rare or altogether absent (identifying 45/63 sites and 31/43 sites for the

neutral and selective models, respectively). Equally important, the neutral model identified 5 out of the 7 sites most frequently mutated prior to protease treatment (35, 37, 62, 63, and 77) and the selective model identified 8 out of 14 sites most frequently mutated after treatment (10, 35, 37, 46, 62, 63, 71, and 77). It was rare for the computational model to predict a site to have low or no tolerance to mutation while the Stanford database indicated otherwise (grey shading

Figure 4.2E, F). However, a few cases of under-predictions of mutational tolerance did occur, most notably at site 93 for the neutral model (predicted 0.8%; database 33.7%) and site 90 for the selective model (predicted 0.01%; database 30.7%). Both models had a weak tendency to over-predict mutational frequencies when compared to the available sequence mutation database (brown shading,

Figure 4.2E, F), especially at the beta-sheet pairing of the dimer interface and dimer flaps. In the case of the neutral model, some of these mutations appear to be at least marginally tolerated by the protease fold, as they become more frequently observed after protease inhibitor treatment (such as at sites 30, 74, 45-47, 55, and 88). Likewise, many mutations at sites that remain "over-predictions" even in the selective model did occur, though at a very low frequency, within the Stanford HIV sequence database (data not shown). We discuss over-predictions of the model further in detail below.

4.5. Evaluation of prediction of specific amino acid types tolerated at each site

Having shown that each model is able to successfully discriminate the overall tolerance to mutation at many HIV protease sites, we next evaluated whether the models

were able to correctly capture the specific amino acid types observed in the Stanford database. We focused on evaluating the performance of the selective pressure model, as the mutations observed in the absence of protease are primarily a smaller subset of the mutations observed and predicted under selective pressure (data not shown).

Table 4.1 gives a complete listing of mutations predicted by the selective model and compares the predicted frequencies to those observed in the HIV Stanford database after protease inhibitor treatment. Mis-sense mutations experimentally found in to display a wild-type phenotype for peptide cleavage and HIV viral replication are also shown [106]. This study evaluated roughly 50% of all mutations reachable by a single nucleotide change from the HIV-1 protease consensus sequence. Model predictions matching amino acid types observed in the Stanford database sequences or the set of experimentally characterized point-mutants displaying wild type behavior are shown in bold, red typeface. Residue types unlikely to be predicted by the selective model, as they are greater than one nucleotide mutation away from the native residue type ($W_{SUM} = 0$), are shown in blue.

	Selective Model		Stanford Database		Mis-Sense Mutagenesis
	% Non-Native	Predicted Mutations	% Non-Native	Observed Mutations	Observed Mutations
1P	15.8	Q[11] R[2] T[1] L[1] S[1]	0	--	L H
2Q	1.6	K[1] E[1]	0.1	--	E
3I	4.0	V[3] L[1]	0.1	--	L N
4T	1.2	S[1]	0.1	--	S
5L	0.2	--	0	--	--
6W	80.4	S[45] L[29] R[5]	0.1	--	G L
7Q	0.2	--	0.3	--	H
8R	0.3	--	0	--	--
9P	0	--	0	--	S
10L	21.55	I[11] V[9] M[1] F[1]	42.7	I[33] V[4] F[4] R[1]	I V
11V	0	--	0.8	I[1]	I
12T	7.0	P[4] K[1] R[1]	9.2	S[3] P[2] A[2] I[1] K[1] N[1]	S I
13I	1.9	V[1]	19.3	V[20]	V
14K	16.9	I[6] T[5] R[1] Q[2]	8.6	R[9]	T M N Q
15I	1.6	M[1] V[1]	17.3	V[17]	V
16G	0	--	3.6	E[3] A[1]	--
17G	0	--	0.8	--	--
18Q	3.2	L[2] E[1] K[1]	1.6	H[1]	L H R
19L	17.0	V[7] I[4] F[3] Q[1] R[1] P[1]	9.8	I[7] V[1] Q[1] T[1]	T
20K	6.7	R[3] Q[3] T[1]	18.4	R[8] I[5] T[3] M[2]	M
21E	9.0	V[4] Q[3] K[1]	0	--	Q V K
22A	3	S[3]	0.4	--	--
23L	0.7	--	0.8	I[1]	--
24L	8.4	I[8] V[1]	5.1	I[5]	--
25D	NA	NA	NA	NA	NA
26T	0.6	S[1]	0	--	--
27G	0	--	0	--	--
28A	3.6	S[4]	0	--	--
29D	0	--	0	--	--
30D	15.8	E[6] H[5] N[4]	12.0	N[12]	E
31T	0	--	0	--	--
32V	21.3	I[21]	3.7	I[4]	L
33L	0.5	--	7	F[4] I[1] V[1]	V
34E	12.2	K[7] D[2] Q[2] G[1]	1.4	Q[1]	A
35E	35.6	D[21] G[12] Q[2]	29.5	D[30] G[1]	S
36M	5.0	I[3] T[2] V[1]	28.7	I[26] V[2] L[1]	I
37N	30.6	D[30]	32.8	D[14] S[11] E[3] T[3] H[1] A[1]	--
38L	0.8	I[1]	0	--	--
39P	3.6	T[3] S[1]	2.0	S[1] Q[1]	T
40G	0	--	0	--	--
41R	7.2	K[4] G[2]	18.9	K[19]	K I
42W	0.1	--	0	--	G
43K	37.0	T[21] Q[12] E[2] N[2] R[1]	2.74	T[3] R[1]	N Q
44P	0.3	--	0	--	T R
45K	17.8	R[11] T[5] Q[2]	2.4	R[2]	R T Q
46M	98.2	I[42] L[30] V[24] T[2] K[1]	27.5	I[18] L[9]	L
47I	100	T[82] V[18]	1.7	V[2]	--
48G	0	--	3.8	V[4]	S H
49G	0	--	0	--	--
50I	4.8	L[3] V[1]	1.5	V[1] L[1]	L

	Selective Model		Stanford Database		Mis-Sense Mutagenesis
	% Non-Native	Predicted Mutations	% Non-Native	Observed Mutations	Observed Mutations
51G	0	--	0.1	--	--
52G	0	--	0	--	--
53F	3.9	Y[4]	3.9	L[4]	Y I L V
54I	1.1	V[1]	21.9	V[18] L[2] M[1] T[1] A[1]	L
55K	22	T[9] N[5] I[4] Q[2] R[1] E[1]	2.5	R[3]	R T I N Q
56V	0.6	F[1]	0	--	--
57R	0.5	K[1]	9.2	K[9]	K
58Q	18.0	K[8] E[7] R[3]	3.4	E[3]	E
59Y	0	--	0	--	--
60D	8.9	N[7] E[1]	8.7	E[9]	E
61Q	4.2	K[1] R[1] E[1]	4	E[3] H[1] N[1]	R
62I	20.3	V[20]	29.7	V[30]	L
63L	88.2	P[83] Q[2] I[1] V[1]	85.1	P[72] S[4] A[4] T[3] Q[2] H[1]	I V P R
64I	3.0	V[3]	22.9	V[20] L[2] M[1]	L
65E	26.1	V[14] D[11] K[1] Q[1]	1.7	D[2]	--
66I	0.5	V[1]	1.5	F[1]	L V F
67C	NA	NA	NA	NA	NA
68G	0.1	--	0.3	--	--
69H	59.1	Y[20] Q[18] L[13] D[3] R[3] N[1]	7.5	K[3] Q[2] Y[1] R[1] N[1]	R L Y Q
70K	9.0	R[4] T[4] Q[1] N[1]	3.2	R[3]	T N
71A	81.8	T[55] V[24] S[2]	38.9	V[27] T[10] I[2]	L
72I	79.4	R[48] K[21] S[4] V[3] N[2] M[1] T[1]	14.6	V[8] T[3] M[1] L[1] E[1]	T L V
73G	0	--	10.3	S[8] T[2]	--
74T	4.7	K[2] S[1] R[1]	6	S[4] A[1] P[1]	S
75V	0.3	--	0.7	I[1]	--
76L	1.4	F[1]	1.4	V[1]	--
77V	28.3	F[24] I[4]	32	I[32]	--
78G	0	--	0	--	--
79P	0	--	0.7	--	--
80T	13.5	S[13]	0	--	--
81P	0.5	--	0	--	--
82V	7.1	A[2] M[2] E[1] I[1] G[1]	26.4	A[20] T[3] I[2] F[1] S[1]	I L T
83N	2.5	D[1] K[1]	0.2	--	--
84I	3.1	V[2] L[1]	11.3	V[11]	--
85I	5.2	V[4] L[1]	2.8	V[3]	--
86G	0	--	0	--	--
87R	0	--	0	--	--
88N	9.2	K[6] D[2] S[1]	9.4	D[7] S[2]	--
89L	0	--	2.6	M[1] V[1]	V
90L	0	--	30.7	M[31]	--
91T	4.3	S[2] A[1] R[1]	0.1	--	A N
92Q	3.2	L[1] E[1] R[1]	2.5	K[1] R[1]	L
93I	7.5	L[7] T[1]	33.7	L[34]	M V F
94G	0	--	0	--	--
95C	NA	NA	0	NA	NA
96T	0.4	--	0	--	--
97L	0.3	--	0	--	--
98N	61.3	T[48] I[5] S[5] D[2] K[1]	0	--	S
99F	0.4	--	0.1	--	L

Table 4.1 Comparison of computationally predicted HIV-1 protease mutations to clinical mutations and mutations with a wild-type phenotype after mis-sense mutagenesis

The subtype-B consensus amino acid type for each of the 99 HIV-1 protease sequence positions is listed in column 1. Non-consensus amino acid types computationally predicted to be tolerated by the selective model are given in column 3. Superscripts denote predicted frequencies of occurrence.. Column 2 gives the predicted frequency with which the native amino acid will mutate at each site (1-probability(native)). Column 4 lists the predicted mutational frequency of each site, as given by the Stanford database (see **Section 4.3.1**). The frequency of occurrence of all mutations found at each site within the Stanford database is given in Column 5. Only amino acids predicted by the model or observed within the database at a frequency greater than 0.5% are shown. Column 6 lists amino acid types experimentally observed to display the wild-type phenotype at each position, as taken from [106]. This study tested approximately half of all possible mis-sense mutations. Red coloring denotes a computationally predicted amino acid type

that was also observed either in the database or the mis-sense mutagenesis study. Blue coloring denotes amino acid types unlikely to be predicted computationally as they require two mutations at the DNA level.

In contrast to other studies that focus on the substrate and inhibitor binding sites to model emergence of HIV-1 resistance mutations [100, 107], the selective model we present here was able to predict a wide variety of specific amino acid types observed at all sequence sites, regardless of structural location. At many positions distant from the substrate binding site (10-20, 31, 33-45, 55-79, 83, 85, and 88-89) the amino acid types selected by the model showed strong agreement with the observed sequence tolerances. Sites such as 11V, 17G, 31T, 38L, 40G, 42W, 44P, 56V, 59Y, 68G, 75V, 78G, and 79P were correctly predicted to be intolerant or barely tolerant to mutation while other sites such as 10L, 12T, 14K, 19L, 20K, 35E, 36M, 37N, 39P, 43K, 45K, 55K, 63L, 69H, 70K, 71A, 85I and 88N had strong overlaps in the predicted and observed amino acid types. (Note that we denote sites by the sequence position followed by the one-letter-code of the amino acid residue present in the consensus sequence). Model performance at sites near the substrate-binding site (23-30, 32, 80-82, 84, and 86-87) and dimer flaps (46-54) was also strong. Sites of low or no tolerance to mutation were correctly identified both within the binding site (23L, 26T, 27G, 29D, 81P, 86G, 87R) as well as within the glycine rich dimer flaps forming the upper portion of substrate binding cavity (49G, 51G, 52G; 48G was incorrectly predicted to be intolerant to mutation by the model). Sites frequently shown to mutate after inhibitor treatment either within the substrate-binding site (24L, 30D, 32V, 82V, and 84I) or dimer flaps (46M, 47I, 50I, 53F, and 54I) were also identified with high fidelity. For each of these sites, the model was able to correctly predict all specific amino acid types observed in the mutational database, with a few exceptions at sites 53, 54, and 82.

As noted above, the model performed less well for several sites within the beta sheet pairing of the dimer interface (1-9 and 90-99). Database sequences showed almost no tolerance to mutation at these sites, with exceptions at sites 90 and 93. In contrast, the model predicted a moderate or high tolerance to mutation at numerous dimer interface sites (2Q, 3I, 4T, 91T, and 92Q, showed moderate to small over-predictions while 1P, 6W and 98N showed large over-predictions) while under-predicting tolerance to mutation at sites 90 and 93. In some cases, amino acids at the dimer interface predicted by the model but not observed in the Stanford database sequences were functionally tolerated when sampled during the in vivo screen.

With the exceptions noted above for the beta-sheet dimer interface sites, the overall frequencies of mutation predicted by the model matched the mutational frequencies observed within the Stanford database fairly well. The model correctly predicted every mutation observed in 3% or more of the database sequences, except in nine cases. Mutations T12S, L33F, N37S, F53L, and L90M were predicted by the model at a very low frequency (0.01%-0.03%; Table 4.1 shows only predicted and observed frequencies >=0.5%) and mutations G16E, K20I, G48V, and G73S were not tolerated at all by the model. Other under predictions of mutational tolerance occurred for four mutations distal to the binding site (I13V, I15V, M36I, and I64V) and two positions near the substrate binding site or dimer flaps (I54V and V82A), where the model predicted mutations at much lower frequencies (1%-3%) than observed within the database sequences (17%-26%). A possible explanation for the low predicted frequencies at these sites may be that the experimentally observed mutations are highly destabilizing or

require the presence of additional compensatory mutations, whereas our model only considers single, independent mutations

As discussed above, over-predictions of tolerance to mutations by the model with respect to the Stanford database sequences also occurred at several sites. The majority of such over-predictions occurred for solvent exposed residues (21E, 43K, 45K, 55K, 65E, 69H, and 72I). At many of these sites, amino acids predicted by the model were found within the experimental mis-sense mutagenesis data, suggesting that the model might be predicting mutations functionally tolerated by HIV protease but not observed at high rates clinically. Over-prediction at other structural sites was rare, but did occur for two positions contacting the substrate-binding site (32V, 47I), one position in the dimer flap (46M) and one core residue (71A). Finally, we note that some computationally predicted mutations not currently observed with the Stanford sequence database might be yet to be recognized resistance mutations. At least three mutations (M46V, F53Y, and N83D) predicted by the selective computational model, but not appearing in the database sequences at appreciable frequencies, have now been identified as HIV drug resistant mutations[108, 109].

4.6. Evaluation of Overall Model Performance

In order to quantify the overall performance of the selective model using a standard metric, we calculated true positive (TPR) and false positive (FPR) rate of identifying HIV Stanford database mutations occurring at three different frequencies (Figure 4.3, black curves; grey curves will be discussed below). Each of 1,746 possible mutations (18 amino acid types, excluding the native amino acid residue and cysteine,

allowable at 97 sites) was computationally classified as either tolerated or forbidden based on whether their predicted frequency was greater than a given computational threshold (see **Section 4.3.7**). These predictions were then compared to the set of experimental frequencies, and receiver-operator characteristic (ROC) curves were constructed.

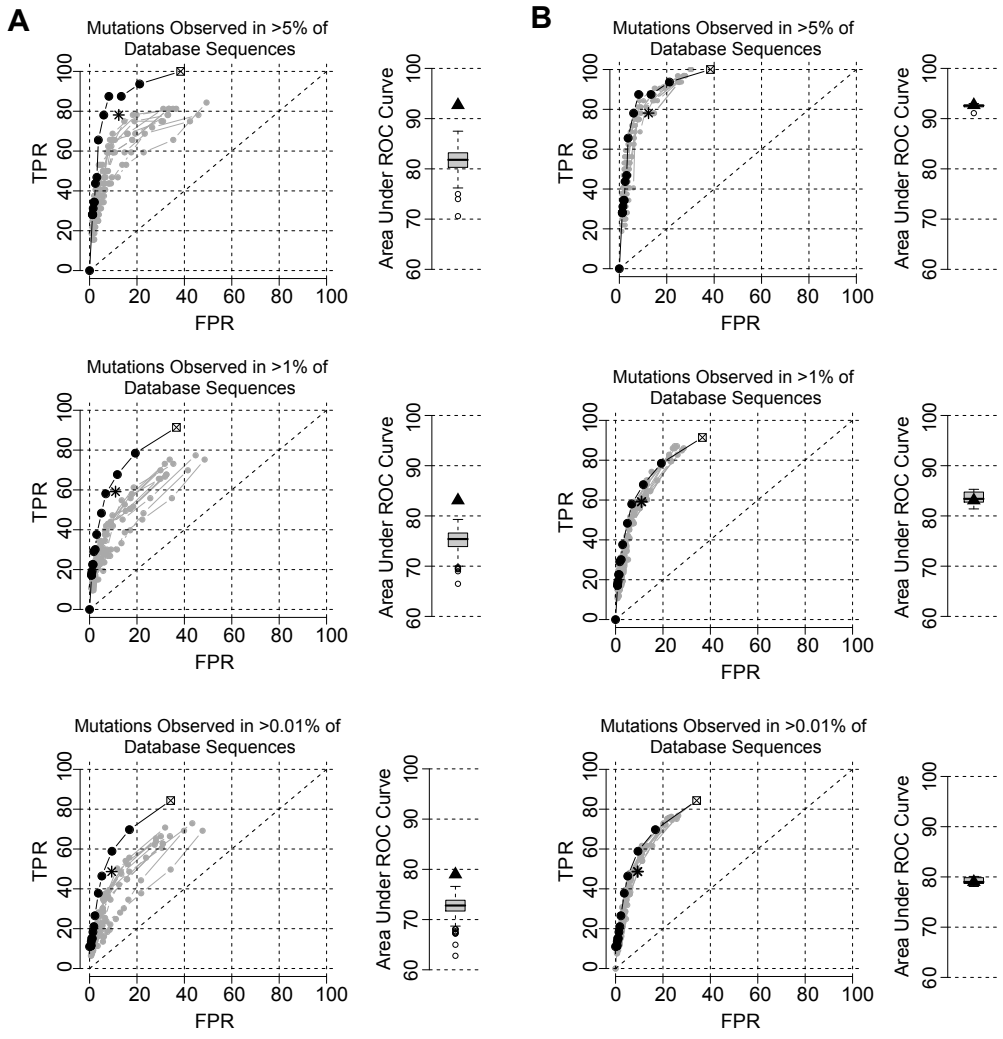


Figure 4.3 Model ROC curves and AUC values.
Receiver operator curves (ROC) were constructed to determine the overall model performance in identifying mutations observed within the HIV-1 Stanford database sequences above three thresholds. ROC curves for predictions of mutational tolerance generated by calculating optimal ERES scores calculated from a crystallographic ensemble of structures are depicted by black circles in all panels in **(A)** and **(B)**. Panel **(A)** also shows 9 representative ROC cures, depicted by grey circles, for predictions of mutational

tolerance made using ERES scores taken from fixed backbone structures crystallized in the absence of mutation. Each of these fixed backbone structures was used as a starting template to computationally generate a "backrub" ensemble of structures as described in Section 4.3.4. Panel **(B)** shows the resulting ROC curves when predictions of mutational tolerance were calculated using the optimal ERES scores observed among each of 9 representative "backrub" ensembles (grey circles). Area under the ROC curves (AUC) are shown for predictions made using the full crystallographic ensemble (black triangles in **(A)** and **(B)**), predictions made independently using each of the 263 crystal structures as a single fixed backbone template (grey box-plots in **(A)**), and predictions made using 11 backrub ensembles, each generated from a different structure originally crystallized in the absence of mutation (grey box-plots in **(B)**). The two null models described in the text (single mutation at the nucleotide level and amino acid types chemically similar to native) are denoted by a square and asterisks, respectively.

The selective model was very successful at identifying mutations appearing in over 5% in the inhibitor treated database sequences, finding over 85% of mutations correctly with a FPR of under 10%. The model performed somewhat less well at predicting mutations occurring in >1% or >0.01% of sequences within the database, but still identified around 80% of known mutations, with an error rate just under 20%. For comparison, we also calculated TPR and FPR for two simplified null models by selecting amino acids at each sequence site that (i) were chemically similar to native or that (ii) were accessible by a single base change at the nucleotide mutation level from the native (subtype B consensus) sequence (asterisks and squares, respectively, in Figure 4.3; see also Supplementary Table C.4). The null model consisting of selecting only amino acids chemically similar to native eliminated a much larger fraction of the total possible mutational space than the null model of nucleotide substitution. This resulted in the chemically similar null model incorrectly eliminating many mutations observed within the database sequences, and resulted in a TPR ~20-30% less than the null model of nucleotide substitution. On the other hand, while the set of amino acids one mutation away from native at the nucleotide level contained the more of the experimentally observed mutations, it also contained ~30% more false positives than seen when taking only amino acid types chemically similar to native. In contrast, the selective

computational model, which uses estimates of mutational effects on stability and function to select tolerated amino acid types, selects less false positives (~20% less) than the null nucleotide mutational model while, at the same time, selecting more true positives (~10-20% more) than a null model of chemically similar amino acid types.

4.7. Prediction of known major and minor drug resistance mutations (DRMs)

We next examined in detail the trade-offs in fold, dimer, and substrate stability predicted by the model for several known major and minor drug resistance mutations (DRMs) [110]. Major and minor DRMs typically both show an increased frequency of mutation after protease inhibitor treatment, while only major DRMs are defined as those that have been directly implicated in inhibitor resistance. Figure 4.4 shows mutations identified by [110] to be major (black squares, Figure 4.4D) or minor (grey squares, Figure 4.4D) DRMs with respect to 8 protease inhibitors used in the clinic (data for all drugs except Nelfinavir (NFV) are considered in the context of combination with a 9th protease inhibitor, Ritonavair (RTV)). For each major and/or minor DRM, the predicted change in $\overline{ERES}$ score, relative to the native residue type, for fold stability (Figure 4.4A), dimer stability (Figure 4.4B), and peptide binding affinity predicted by the model (Figure 4.4C) is shown. With the exception of a small number of DRMs strongly disfavored by the model for requiring more than one nucleotide mutation from native (blue labels, 3 major DRMs and 6 minor DRMs), the model was able to predict mutational tolerance for 80% of the known minor DRMs and major DRMs (16/20 and

34/42, for major and minor DRMs, respectively; red labels, predicted tolerated; black labels, predicted not tolerated).

Most major DRMs were predicted to have a large, destabilizing effect on at least one of the 10 tested endogenous peptides ($ERES_{PEPTIDE}$ score changes are shown as the change in the sum of scores for all 10 peptides before and after introducing the mutation). This suggests that at least in some cases DRMs may be able to appear and cause disruptive changes even within the "substrate envelope" (put forth in [111] and [112] as the consensus volume occupied by the set of natural substrates). It has been shown that HIV-1 protease peptide cleavage sequences can, in some cases, co-evolve along with the appearance of DRMs, and this mechanism may allow for some compensation of this predicted peptide destabilization[113, 114] [115-117]. In addition, a few major DRMs predicted to have destabilized peptide affinity were also predicted to have additional negative effects on fold and/or dimer stability. Mutations at sites 47 and 50 were predicted to greatly destabilize the dimer interface, while other mutations were predicted to either moderately (82F/L/I/A, 47V, 84V) or strongly (48V, 47A and 53L) destabilize the protease fold. At least some of these predicted destabilizing effects may require the presence of minor DRMs or other stabilizing mutations in order to maintain native viral efficiency.

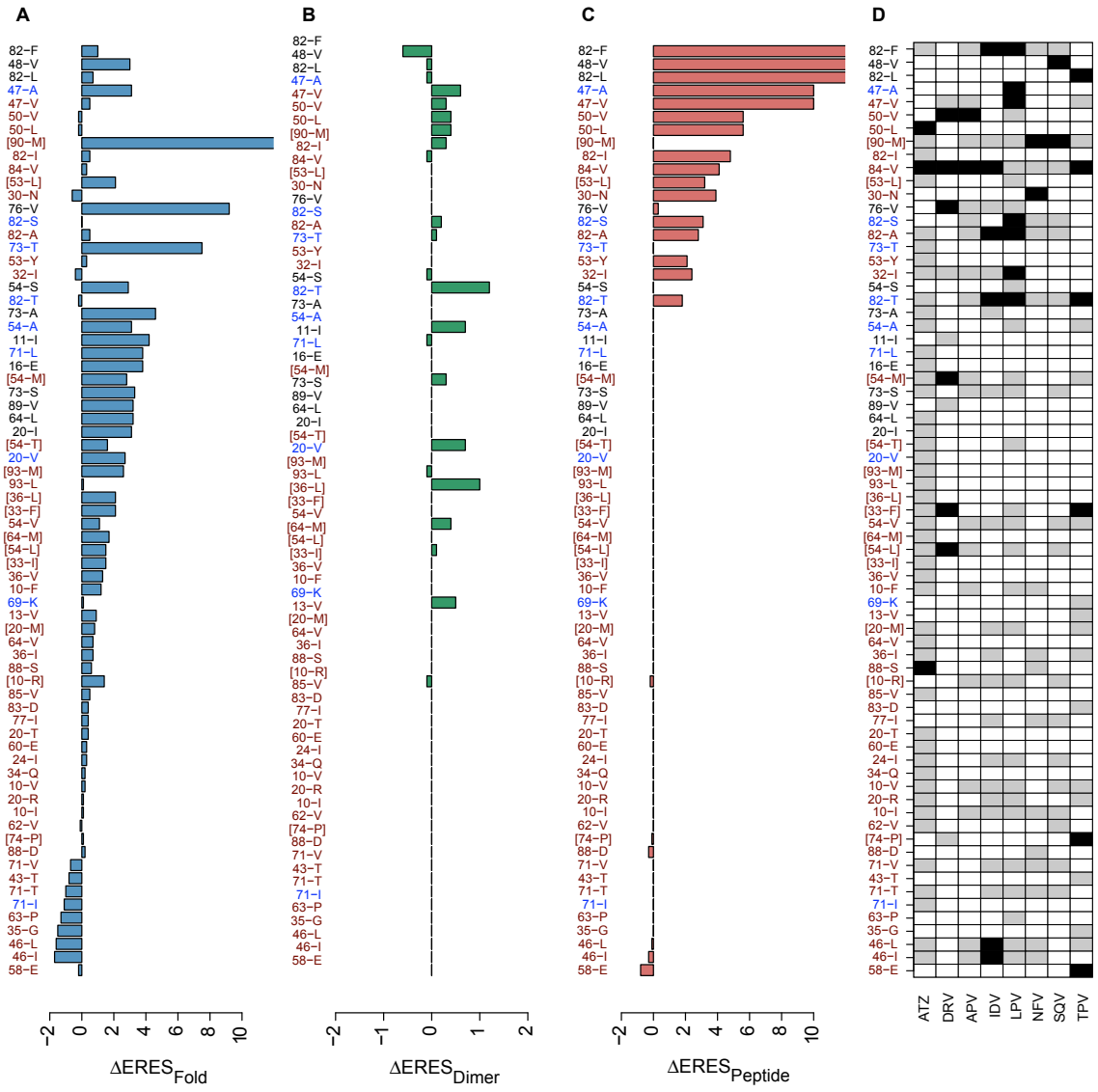


Figure 4.4 Model prediction of major and minor DRMs.
Calculated changes in $ERES_{FOLD}$, $ERES_{DIMER}$, and $ERES_{PEPTIDE}$ scores (relative to that native, consensus subtype B) are given in **(A)**, **(B)**, and **(C)** for each drug resistance mutation identified by REF. Mutations identified as major (black squares in **(D)**) or minor (grey squares in **(D)**) drugs resistance mutations are shown with respect to each of 8 drugs (ATZ, Atazanavir + Ritonavir; DRV, Darunavir + ritonavir; APV, Fosamprenair + Ritonavir; IDV, Indinavir + Ritonavir; LPV, Lopinavir + Ritonavir; NFV, Nelfinavir; SQV, Saquinavir + Ritonavir; TPV, Tipranavir + Ritonavir). $ERES_{Peptide}$ score changes are shown as the change in the sum of scores for all 10 peptides before and after introducing the mutation. Note that reliable $ERES_{PEPTIDE}$ scores could not be calculated for mutations at residue 47 due to a large steric clash between a modeled peptide and the native residue at that site. Blue labels denote mutations not predicted as tolerated by the selective computational model, as the required two DNA mutations from native.

For a few major DRMs (90M, 76V, 54M/L and 33F) there were no predicted destabilizing effects at the substrate-binding interface. Instead each of these mutations was predicted to be strongly destabilizing to protease fold stability, and in two cases (90M, 54M) also predicted to be destabilizing to the dimer interface. This suggests that larger scale structural changes or the presences of compensatory co-mutations might also be necessary in order for the HIV protease fold to accommodate these major DRMs. Interestingly, there were also a few DRMs at sites which neither contact the peptide substrates nor form the dimer interface. These DRMs were predicted to either stabilize (46L, 46I, 58E) fold or peptide stability or be mostly neutral (88S, 74P) with respect to all structural and functional constraints considered. These predicted neutral mutations may play a role correlated with other DRMs in affecting fold, dimer or peptide stability. Moreover, the predicted stabilizing mutations may be important in alleviating the negative effects of other simultaneous sequence changes.

For the minor DRMs, we observed a different pattern of stability effects. Approximately half of minor DRMs were predicted to have large to moderate destabilizing effects on fold stability (such as 36I/L/V, 11I, 16E, 33F, 54V/T), while the other half were predicted neutral or stabilizing with respect to fold stability (such as 83D, 77I, 20T, 20R, 62V, 71I/T/V). The predicted destabilizing effects of at least a few minor DRMs could be relieved by the presence of additional compensatory (or correlated) mutations while the stabilizing minor DRMs could themselves play a compensatory or complementary role to the destabilization seen in the major DRMs. Of the minor DRMs, mutations at site 54 and 93 were predicted to destabilize the dimer stability. Whether this dimer destabilization would be able to be compensated for by additional accumulation of

mutations, or whether the destabilization itself is playing a functional role in drug resistance is unclear.

4.8. Importance of backbone flexibility: Crystallographic versus computationally generated conformational ensembles

Finally, we compared predictions made using multiple structural backbone templates (representing various conformational "snapshots" of HIV-1 protease backbone flexibility), to predictions made by selecting just one HIV protease backbone to use as a template for the calculations. To test this, we compared predictions of mutational tolerance made using $ERES_{FOLD}$ and $ERES_{DIMER}$ scores obtained from a single crystal structure (fixed backbone) with the previously presented model results which had calculated optimal $\overline{ERES}_{FOLD}$ and $\overline{ERES}_{DIMER}$ scores over a set of crystallographic ensemble members (identical $\overline{ERES}_{PEPTIDE}$ scores and parameters were used in all calculations). As noted above, many of the HIV-1 protease ensemble members originally contained a small number of non-native substitutions. Thus it is possible that at least some of the variation observed among the crystallographic ensemble members might be due to structural changes and/or accommodations directly resulting from the original presence of non-native substitutions. To remove any such "memory" of mutations originally present at crystallization, we examined fixed backbone predictions made on each of 11 structures originally crystallized with the native subtype B sequence

Independent of which single crystallographic structure was used for the calculations, we found ROC curves from each fixed backbone model to perform much worse than the original model, which incorporated structural information over an

ensemble of backbones (Figure 4.3A, grey lines show 9 representative ROC curves, each generated using a single crystallographic HIV-1 protease structure). Comparison of the area under the ROC curves (AUC) generated by every possible single backbone computational model (Figure 4.3A, grey box plots, AUC calculated for predictions made on each of the 263 crystallographic ensemble structures as a single fixed backbone) to the AUC obtained from the original ensemble model (Figure 4.3A, black triangle), showed that predictions made using a single fixed backbone structure were in no case able to match the performance seen by using the entire ensemble of structures.

We then asked whether, starting from a single fixed backbone structure, we could computationally generate an ensemble of structures that would convey structural information similar to that seen in the original crystallographic structural ensemble. If true, this would enhance the general applicability of the model presented here for predicting mutational tolerance in other protein targets for which there might not be as large an ensemble of crystal structures available. Using each of the 11 HIV-1 protease structures crystallized in the absence of mutation as independent starting templates, we computationally generated ensembles of "near-native" backbone conformations by using a backbone flexibility move termed "backrub" [33] (see Section 3.3.3; the RMSD of each ensemble member to the original starting template varied, but was less than 0.5 Angstroms for most backbone conformations, and never exceeded 1.4 Angstroms for any ensemble member; for RMS fluctuations see Supplementary Figure C.5). Each computational ensemble was used to model all single-point mutations as described above for the crystallographic ensemble and the lowest $ERES_{FOLD}$ and $ERES_{DIMER}$ scores seen among the backrub ensemble members were selected and used in Eq(1) and Eq(2) to

predict mutational tolerance. Remarkably, independent of which starting fixed backbone structure had been used, model predictions made on any of the 11 computationally generated ensembles resulted in ROC curves and AUC values virtually indistinguishable from predictions made using the original crystallographic ensemble (Figure 4.3B).

4.9. Structural and energetic implications for representative mutations

In order to gain insight into how structural flexibility might improve predictions of HIV-1 protease mutational tolerance, we examined model predictions at several sites in greater detail. We focused on several mutations that had been originally present in at least one of the 263 HIV-1 protease crystal structures (all such sites, including the number of crystal structures originally containing each specific mutation, are given in Supplementary Figure C.3 and Supplementary Figure C.4). Predictions made using the full crystallographic ensemble, which included one or more structures containing a backbone conformation originally crystallized with the amino acid substitution under consideration) were then compared to predictions made using single fixed backbone structures crystallized in the absence of any mutations.

Figure 4.5A shows two mutations (I62V and M46I) for which predictions made on a crystallographic ensemble, single fixed backbone, or backrub generated ensemble were all equally accurate in predicting the frequency with which a mutation was observed in the HIV Stanford database (red bars). At such sites the predicted frequency of mutation made using the full crystallographic ensemble (black bars) was similar to the median prediction resulting from using any one of 11 single fixed backbone structures (grey

bars). Further, computational introducing flexibility by generating a backrub ensemble from each of the fixed backbone structures (grey striped bars) resulted in virtually identical predictions.

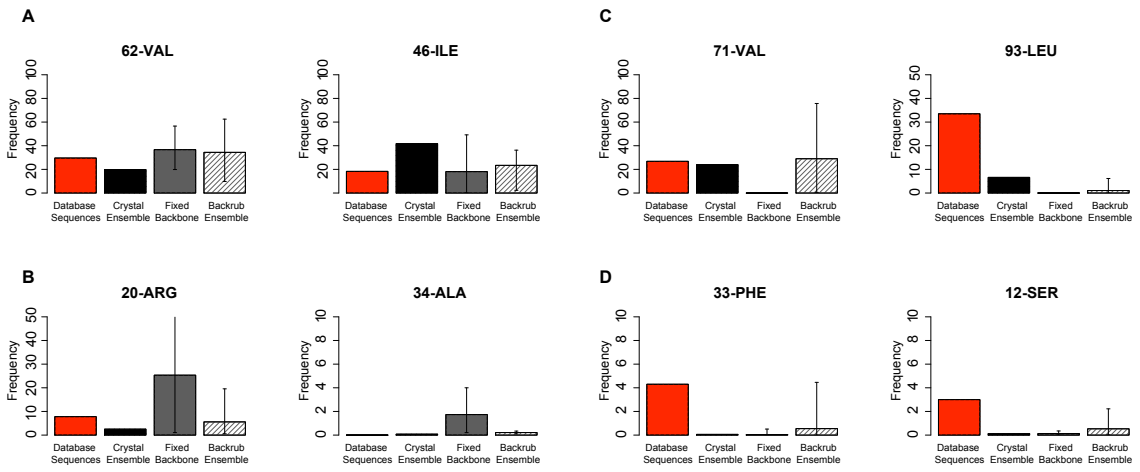


Figure 4.5 Comparison of predicted frequencies with and without backbone flexibility.
The frequency with which 8 mutations were predicted when using either an ensemble of backbones or a single fixed backbone structure are compared with the frequency with which each mutation was observed to occur in patients treated with HIV-1 protease inhibitors in the Stanford database (red bars). Black bars denote predicted mutational frequencies observed when using optimal ERES scores calculated using the full ensemble of 263 crystal structures. Grey bars give the mean mutational frequencies observed when ERES scores were taken from 11 fixed backbone structures, crystallized in the absence of mutation. Grey striped bars depict the mean mutational frequencies observed after each of the 11 fixed backbone structures was used to computationally generate an ensemble of "backrub" structures, and these ensembles were used to predict optimal ERES values. Vertical bars denote the maximum and minimum mutational frequency observed.

At other sites, we observed that incorporation of flexibility (*via* using either a crystallographic or backrub ensemble) resulted in mutational frequencies more inline with frequencies observed in the Stanford database (Figure 4.5B). In many such cases, we observed ERES scores calculated on single fixed backbones were often fairly "flat" (e.g. similar for all 19 amino acid types) in comparison to $\overline{ERES}$ scores calculated over crystallographic or computational ensembles of backbones (data not shown). This

phenomenon was likely responsible for the significant reduction in the FPR rate observed for backrub ensembles as compared to fixed backbone in Figure 4.3. (Note that in a few cases the FPRs for fixed backbone models were actually greater than the null model of nucleotide substitution; this indicates that fixed backbone predictions at some sites had ERES values so flat that amino acid types with W_{SUM} values set to zero were occasionally selected as tolerated).

In a few cases, flexibility appeared to be crucial for correctly predicting tolerance to mutations observed within the Stanford database. Figure 4.6 shows two cases where large to moderate clashes are observed upon modeling a mutation onto fixed backbones structures (top row; 71V, left-side panel and 93L, right-side panel). In each case, the clashes were resolved when the mutation was modeled onto a backbone computationally generated from each fixed backbone structure (middle row). Further, the mutations modeled onto the computationally generated backbones had structures and $ERES_{FOLD}$ scores comparable to those seen in crystallographic structures, which had originally contained the mutation (bottom row; $ERES_{DIMER}$ and $ERES_{PEPTIDE}$ scores not shown). Figure 4.5C confirms that the mutations 71V and 93L were predicted to be tolerated when modeled onto either crystallographic or backrub ensembles, but never when modeled onto a single fixed backbone crystallized in the absence of mutation. Other mutations for which we observed a similar trend included 24I and 77I.

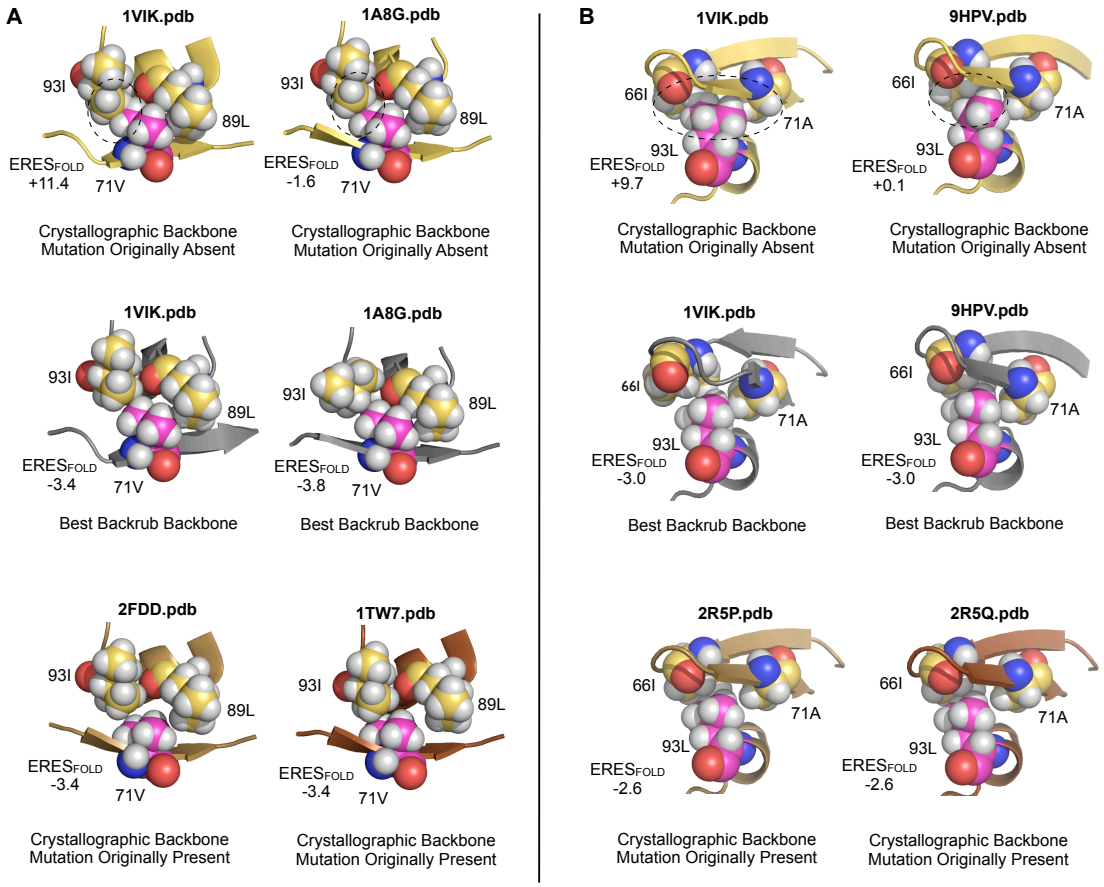


Figure 4.6 Example A71V and M93L structural models.
Mutations A71V (left) and M93L (right) result in large to moderate steric clashes when modeled onto two fixed backbone structures, crystallized in the absence of any mutations (top row). Each fixed backbone structure was used as a starting template to computationally generate an ensemble of backrub backbones. The optimal "backrub" backbone (e.g. the one with the lowest calculated $ERES_{FOLD}$ score), alone with its calculated $ERES_{FOLD}$ score, selected when using each of two fixed backbone starting structures is shown for each mutation (middle row). For comparison, $ERES_{FOLD}$ scores and structures are shown for two members within the crystallographic ensemble that had originally contained A71V or M93L at the time of crystallization (bottom row).

Finally we noted that several mutations within the Stanford database sequences that had been poorly predicted when using the full crystallographic ensemble were observed to be tolerated when using computationally generated ensembles (Figure 4.5D; note 33F and 12S were discussed above as representative "failures" of the selective model). These mutations were also predicted (at much smaller frequencies) by a few

fixed backbone structures. This same trend was further observed in other mutations under-predicted by the crystallographic model including 13V, 15V, 53L and 64V (Supplementary Figure C.3).

4.10. Discussion

We have shown an all-atom, computational model that incorporates structural and functional constraints on mutational tolerance is able to predict a large portion of the clinically observed HIV-1 protease sequence space. We note that we have used a previously published and established energy function and design protocol, and that there was no explicit parameterization with respect to the actual identities of amino acids selected to be tolerant or intolerant to mutation at each protease site. Instead, only the relative frequencies with which the various structural and functional constraints operated (as well as a new solubility parameter $\text{PENALTY}_{\text{POLAR->HP}}$) were optimized in order to best predict overall mutational tolerance (Supplementary Figure C.1).

The model we present makes testable hypotheses about the effects of specific mutations on the stability of HIV-1 protease fold, dimer interface, and substrate binding that, in some cases, can be experimentally verified. For example, the model predicted large substrate destabilization effects for V47A and V82A/F/T and these mutations are known to display increased viral replication in viruses with mutations in either the NC/p1 or p1/p6 cleavage sites [90, 117, 118]. Likely, these cleavage site mutations are relieving destabilizing clashes predicted by the model at the substrate-binding interface. In another example, incorporating A71V (predicted by the model to stabilize the protease fold) into double and triple mutants with reduced replicative ability containing either 36I/54V or

36I/54V/82T (all predicted to be destabilizing) has been shown to improve replication to better than the wild-type virus levels [119]. While the exact magnitude of stabilizing and destabilizing trade-offs predicted by the model for each individual mutation is only an estimate, the patterns of compensatory effects and functional tradeoffs may nevertheless be informative in understanding the numerous sequence mutations observed with HIV-1 protease.

It is important to note that the model presented here attempts to explain the complex mutational patterns observed within clinical HIV-1 protease sequences *only* by considering additive effects of independent, single point mutations. While we have shown this simplifying assumption can still lead, in many cases, to good predictions of HIV-1 protease mutational tolerance, correlations among the appearance of HIV drug resistance mutations have been widely observed [120, 121]. In a few cases, under-predictions in the model may be due to ignoring non-additive effects of correlated mutational changes. For example, the mutations 30N and 88D have been found to co-vary, and while the model predicts mutational tolerance for both mutations, the frequencies predicted are less than seen in the database sequences for each mutation. We are exploring methods to expand the current model by more directly including and modeling non-additive mutational effects.

In several cases, non-active site mutations have been hypothesized to play a role in inducing conformational changes that would directly allow active-site mutations to be structurally tolerated. This idea could also be easily explored by using an iterated version of the model of computational flexibility already presented. An initial mutation could be computationally introduced and this mutated structure could be used as a structural

template for generating an ensemble of "backrub" backbones. ERES scores for a secondary mutation on this ensemble could then be compared to the ERES scores observed when scored on an ensemble generated the absence of any initial mutation.

One of the strongest conclusions from the work we present here is the importance of incorporating backbone flexibility in predicting the mutational tolerance of HIV-1 protease. Structural changes to accommodate mutations have been noted in several crystallographic studies [122]. Flexibility has also been shown to be important in accurately predicting substrate docking and protein-inhibitor binding energies of HIV-1 point mutants, and changes in protein dynamics after mutation has been widely observed in several molecular dynamics studies [123, 124]. We provide evidence that the protocol presented here for computationally generating backbone flexibility may be able to directly mimic structural information contained within a set a diverse crystallographic structures, as predictions of HIV-1 mutational tolerance were similar regardless of whether an ensemble consisting of crystallographic or computationally generated structures was used. This was surprising, as the conformational variation included within the ensemble of crystallographic backbones included changes induced by substrate and inhibitor binding, as well as structural changes induced by over 65 mutations, some crystallized as point mutations and others in the context of multiple mutations. Use of crystallographic ensembles to model protein conformational flexibility has been described by Vendruscolo and coworkers and shown to be consistent with protein dynamics detected using nuclear magnetic resonance [125]. In this work B-factors calculated over the computationally generated "backrub" ensembles showed a striking similarity to B-factors calculated over the ensemble of crystallographic structures

(Supplementary Figure C.5), as well as to B-factors observed in previous molecular dynamics studies (data not shown, see for example [126]). Thus it is plausible that "backrub" ensembles are sampling a significant portion of dynamic conformations accessible to HIV-1 protease. Finally we note that previous studies on a variety of other protein systems have found computationally generated "backrub" ensembles to improve predictions of side chain dynamics [78], conformations of single point mutations [33], and sequence diversity at protein-protein interfaces [2].

The approach we present here differs from numerous other studies that have studied the structural [127, 128] and functional effects [105, 129-131] of HIV-1 protease mutation on inhibitor binding, by instead making predictions of the mutational effects on the HIV-1 protease structure and function in the absence of any bound inhibitor structure. We imagine the two approaches may be complementary. At least one study so far has shown incorporation of site-specific sequence variability (determined from sequence conservation) may improve the ability of computation to correctly predict drug resistance mutations over consideration of changes in inhibitor binding energies alone [132].

This work also addresses HIV-1 protease substrate recognition and binding from an approach that may complement other studies predicting plasticity among the peptide sequences recognized and cleaved by the native HIV-1 protease sequence [133, 134]. Several non-native mutations predicted by our model are predicted to have significant destabilizing effects on substrate binding. Predictions of mutational changes to substrate peptide sequences that could offset predicted destabilizing effects of non-native HIV-1 protease mutations could prove useful for examining co-evolution between the sequences of HIV-1 protease and its endogenous substrates. Further, overall model performance,

especially at the beta-paired dimer interface, might be improved by the use of any number of established peptide docking or modeling tools. In this work we modeled three HIV-1 protease-bound peptide complexes by using simple sequence threading. Use of more sophisticated protein-substrate modeling techniques could help resolve several large to moderate clashes we observed in our modeled complexes, most notably at site 47, for which the native residue was found to have lowered tolerance or not be tolerated at all.

Finally, while our model does not directly make predictions for improved inhibitors to evade HIV-1 drug resistance, it may complement the strategies put forth by several other groups. Work seeking to develop inhibitors better able to evade resistance mutations, either by designing inhibitors to more strictly encompass the spatial pocket spanned by natural HIV-1 substrates (the substrate envelope hypothesis [112, 135]) or by increasing inhibitor binding affinity to wild type and/or mutant proteases [136, 137], may be strengthened by selectively increasing inhibitor interaction with HIV protease residues intolerant to mutation. The model we present here could complement such an approach, as it successfully predicts HIV-1 protease sites experimentally determined to be structurally and/or functionally intolerant to mutation with high fidelity.

4.11. Conclusion

We have presented a computational model able to predict mutational patterns observed in HIV-1 protease patients, both before and after protease inhibitor treatment. This model makes testable predictions about the structural and functional effects, and possible compensatory trade-offs of specific mutations. Further, it incorporates a computational model of protein flexibility that we show is able to mimic structural

plasticity observed within a large number of crystallographic structures. This model could easily be extended to predict mutational tolerance for any number of other protein systems, but also may offer some unique insights on the problem of predicting HIV-1 drug resistance mutations.

While some drug resistance mutations are shared among the 9 protease inhibitors currently in use, novel drug resistance mutations have appeared with the introduction of each new clinical drug. These mutations could often not be predicted in advance by extrapolating from then currently available inhibitor-protein complexes. The model we present here could provide useful in the prediction of yet undiscovered resistance mutations by suggesting mutations structurally and functionally tolerated by the HIV-1 protease fold and free to contribute to destabilization of new clinical inhibitors. The model predictions could also prove useful for generating a stabilized HIV-1 protease sequence, into which other destabilizing mutations that have the potential to cause loss of inhibitor binding, could be screened. Finally, while we have examined HIV-1 sequence changes predicted starting from the subtype B consensus sequence in this work, the model could easily be extended to using other subtype sequences, and thus could prove useful for exploring the different mutational pathways observed in various HV-1 protease subtypes.

Chapter 5 Conclusion

In this work, I presented a series of computational tools to predict protein sequences compatible with a given three-dimensional protein structure and a set of structural or functional constraints. Section 5.1 describes how these tools have been used thus far and gives a summary of the major findings to date. Section 5.2 discusses several ways in which the work presented here could be expanded and modified for future use in protein sequence prediction and design. Examples of ongoing research projects within the Kortemme lab at UCSF, which directly build upon the computational tools presented in this work, are also given.

5.1. Summary

First, I described how a multi-constraint protein design protocol could be used to design protein interface sequences predicted to bind to multiple interaction partners, and then compared computationally designed interface sequences to naturally occurring amino acid sequences for a number of multi-specific signaling and structural proteins. For proteins with large, flat interfaces that had 3 or more structurally characterized binding partners, the sequences predicted as optimal by the multi-constraint protocol were found to be very native-like, both in terms of amino acid sequence identity as well as in predicted binding affinity of each partner. In contrast, single-state design of the

same interfaces for binding only a single interaction partner at a time (ignoring sequence preferences of all other characterized binding partners) resulted in low native sequence recovery and predicted binding affinities much stronger than calculated for the native interface sequence. Thus the multi-constraint algorithm presented in this work appears to, at least in some cases, successfully integrate or "add up" diverse sequence preferences seen among several characterized binding partners. How the distinct sequence preferences observed among differing binding partners may be exploited in order to re-design protein interfaces or develop small molecules with specific binding preferences will be discussed in the Section 5.2.1. Further, the finding that multi-constraint designed sequences had weaker predicted binding scores overall than single-constraint designed sequences suggests that it might be a general overall trend that multi-specificity may come at a cost of affinity. Thus the transient nature observed in the multiple interactions of many signaling protein interactions may not need to be explicitly considered during sequence design or evolution.

Next, using the hGH-hGHR interface as a model system, I examined whether computation tools originally developed to predict a single, optimal sequence for a set of functional constraints could be further developed to predict sets of amino acids tolerated at each position in a protein-protein interface. I modified the previously presented computational multi-constraint algorithm to predict amino acid tolerances and showed that this modified protocol was able to correctly discriminate several positions within the hGH-hGHR interface that had been experimentally shown to tolerate substitution to a wide variety of differing amino acid types. However, several positions within the hGH-hGHR interface were experimentally found to tolerate only a few non-native amino acids,

and predictions made using a fixed backbone often failed to identify some of these non-native amino acid substitutions as tolerated. Prediction of amino acid tolerance at such "restricted" sites was found to be improved by incorporating backbone flexibility by using a set of computationally generated "near-native" backbones for design. Several structural examples for fixed backbone prediction failures, including sterics and subtle changes in backbone conformation enabling altered hydrogen bonding networks, were then discussed. Finally, based on the use of flexible backbone design, a protocol for automatically discriminating sites with restricted tolerances and outputting a library of computationally designed amino acid choices for each site under consideration was detailed. Section 5.2.2 discusses several ways in which the flexible backbone design protocol presented in this work could be expanded and improved, and Section 5.2.3 outlines how computational design libraries could be developed to "design in" new functionality or enhance existing properties of protein sequences.

Finally, I presented a protocol for predicting single amino acid substitutions tolerated for a protein of great biological relevance, HIV-1 protease. Here, I incorporated multiple structural and functional constraints, by computationally predicting how amino acid substitutions would effect overall protein fold stability, the stability of the dimeric HIV-1 protein interface, and the binding and recognition of each of the 10 known HIV-1 protease substrates. For each site in the HIV-1 protein sequence, a reduced set of amino acid mutations (able to be reached by a single mutation at the nucleotide level) was considered and tolerance to mutation at each site was calculated independently. Even with the simplifications inherent in the model, approximately 80% of amino acid substitutions that occurred in 1% or more of clinical HIV-1 protease sequences were

predicted as tolerated by the structural and functional constraints considered in the computational model. In contrast, the model predicted mutations not yet observed in clinical sequences with a false positive rate of only ~20%. Interestingly, this work also demonstrated that use of a single, fixed backbone as a structural template for design resulted in overall poorer predictive performance than designing on an ensemble of either crystallographically determined or computationally generated backbone structures. Section 5.2.4 discusses some future directions for examining HIV-1 protease mutational tolerance that build upon the tools presented within Chapter 4.

5.2. Future Directions

5.2.1. Multi-State Design Successes: Rewiring Protein Signaling Interaction Networks and Engineering of Conformational Stability

The multi-state design algorithm presented in Chapter 2 was examined in the context of predicting interface sequences optimized for binding to multiple interaction partners, but can easily be generalized to predict sequences optimal *for* or *against* any set of enumerable constraints. Since the work in Chapter 2 was published, the same multi-state design algorithm has been successfully used within the Kortemme lab by Mariana Babor to examine backbone flexibility during the maturation of H3 antibody loops [138] and by Noah Ollikainen to predict mutations in the sequence of the small GTPase Ran to specifically disrupt binding with some interaction partners and not others (data not yet published). These initial successes indicate that the algorithm presented here should prove useful for future design projects, such as designing a single protein sequence for stability in two or more different backbone conformations or incorporating negative design to redesign multi-specific protein interfaces to have specific binding interactions.

Finally, the multi-state design protocol presented in Chapter 2 could easily be modified to perform orthogonal interface design. Here the starting interface sequence between a protein pair, A-B, would be scored and used to estimate the strength of the starting, or wild-type interaction. Both sides of the protein-protein interface could then be designed simultaneously, with each interface sequence design would be scored with respect to its ability to destabilize the wild type-design interfaces A*-B and A-B* while stabilizing the novel orthogonal interface A*-B* (here *'s represent a designed interface sequence for either the protein A or B, while lack of stars indicates the original, native interface sequence). This protocol should be more efficient than previous step-wise methods which require the selection of a mutation(s) on a single side of the protein-protein interface predicted to destabilize the native interaction, followed by computational prediction of mutations on the alternate interface side that would restore, or compensate, the original loss of binding.

5.2.2. Future directions for flexible backbone protein design

Chapters 3 and 4 provided evidence that incorporating signal from design on an ensemble of backbones improves overall prediction of tolerance to amino acid sequence changes. In Chapter 4, evidence was provided that predictions made on a set of computationally derived backbones, generated from a single crystal structure, are comparable to predictions made using a set of experimentally determined backbone structures that have been crystallized under a variety of conditions, including the presence of several known substrates and inhibitors. This surprising finding opens up exciting avenues to continue development of the "backrub" backbone flexibility protocol in order to further improve the accuracy of protein design predictions.

The work presented in this thesis was focused on overall prediction of amino acid tolerances, and as such pre-generation of backbone ensembles for design predictions was found to be necessary in order to perform quick and large-scale calculations. However, in some cases, local and independent optimization of a backbone structure for each specific amino acid change may enhance predictive power. In particular, a procedure, iterating selection of amino acid sequence changes with backrub backbone flexibility moves, could be expected to yield more accurate structural models and Rosetta design scores. Such an iterative protocol, using fragment insertion to incorporate backbone flexibility, was successfully used in the design on the novel protein fold, Top7 [7].

Finally, different magnitudes of protein movement may be appropriate for differing protein design tasks. The redesign of tightly packed protein cores, as well as some sites within a protein-protein interface, may require only the introduction of subtle backbone moves. In contrast, modeling loops or larger scale conformational changes that occur at some protein-protein interfaces could require the generation of larger, more structurally diverse backbone moves. Research is ongoing within the Kortemme lab to explore the utility of varying Monte Carlo simulation temperatures during backrub simulations, as well as incorporating alternative backbone flexibility move-sets such as analytical loop closure, into protein design.

5.2.3. Computational Library Design: Development of an online server for sequence enrichment

In Chapter 3, an algorithm to computationally output libraries of optimal amino acid choices for a set of sequence positions was presented. Since the time of publication of the prediction of hGH-hGHR interface sequence tolerance profiles, the computational

library prediction code has been successfully used within the Kortemme Lab by Mariana Babor to predict antibody-HER2 interface sequence tolerances, and is now being modified by Colin Smith to examine PDZ-peptide interface specificities. Also within the Kortemme lab, Florian Lauck is now incorporating the computational library design protocol into an online prediction server, which will be freely available to the academic community.

This automated library design algorithm could be used to enhance protein design in a number of ways. First, sequence positions predicted to be highly tolerant to substitution with a wide variety of differing amino acid types could be identified and subsets of amino acid types could be selected at such sites to "design in" new protein functionality without disrupting the original protein fold or function. Alternatively, one could focus on sequence positions predicted to be more restricted with respect to tolerance to amino acid substitutions. Amino acids substitutions at such sites would be expected to show strong effects on the original functionality, either by enhancing or completely disrupting the original protein functionality.

While overall strong predictive performance of amino acid tolerances at the hGH-hGHR interface was presented in Chapter 3, several extensions and modifications could be made to the library design algorithm that might improve performance even further. Section 5.2.2 discusses several modifications to the flexible backbone protocol that, if implemented, could be expected to improve library design predictions. Further, the protocol presented in Chapter 3 that examines amino acid tolerances at each interface site independently, can easily be extended to predict correlated mutational sequence changes expected to modify or enhance protein stability or functionality.

5.2.4. Future Directions for prediction of HIV-1 protease mutational tolerance

Chapter 4 discusses a first step towards computationally integrating structural and functional constraints on protein sequence in order to predict amino acid sequence changes compatible with the given constraints. Prediction of specific mutational changes tolerated by a protein fold and function might be particularly useful in anticipating drug resistance escape mutations, or those amino acid sequence changes able to destabilize or weaken protein-drug interaction without destroying original protein functionality. Knowledge of tolerance to specific resistance mutations within protein sequences is often obtained only empirically, after drugs have been developed and administered. New rounds of drug design then seek primarily to avoid already discovered drug resistance mutations, without any a priori knowledge of what unobserved mutations may yet appear. The computational tolerance model outlined in Chapter 4, in combination with library design methodologies developed in Chapter 3, offers the ability to predict libraries of tolerated amino acid sequence variants, focused around a drug or ligand binding site. If necessary, development of a library of variants containing possible resistance mutations could proceed in two stages, with an initial round of sequence variants designed to have enhanced fold or functional stability, followed by a second round of variants allowing slightly more destabilizing sequence changes near putative drug binding sites. Novel drugs could then be screened against such libraries of sequence variants, in order to anticipate each drug's suitability towards maintaining efficacy in the face of possible protein sequence changes.

In addition to developing libraries containing putative drug resistance mutations, the tools outlined in Chapter 4 could be further developed in a number of ways. First, the

model as presented considers only single, independent amino acid changes but could easily be expanded to consider correlated mutational changes by either (1) using a brute force approach to enumerate and score all possible double, and even triple combinations of amino acid changes or by (2) computationally evolving combinations of amino acid changes by ensuring that the sum total of the change in scores for each constraint for all incorporated sequence modifications is within some threshold value. Such a two-pronged approach could allow one to make broad computational predictions about the most likely evolutionary order of accumulation of mutations, as well as specific predictions about the energetic and structural reasons for the appearance of double or triply correlated mutations.

The tools of Chapter 4 could also be modified to examine two other interesting mutational patterns observed for HIV-1 protease. First, HIV-1 protease exists in several differing subtypes, each of which differs in what exact 99-residue amino acid sequence is considered to be "native". The computational model presented in Chapter 4 considers a protein's mutational tolerance to be constrained not only by its amino acid sequence, but also by its underlying DNA sequence, by explicitly dis-favoring amino acid sequence changes that require more than one nucleotide mutation at the DNA level. This opens the possibility that such a computational model might provide useful in predicting and understanding the specific mutational tolerance patterns observed among varying HIV-1 protease subtypes. Finally, it has been observed that the sequence of the peptide cleavage sites, 10 of which are known to be HIV-1 protease substrates, can co-evolve along with the appearance of mutations within the HIV-1 protease binding-site. The computational protocol presented in Chapter 4 could easily be extended to predict tolerance to the co-

evolution of sequence changes occurring in both the HIV-1 protease protein, as well as its substrate peptide.

Bibliography

1. Humphris, E.L. and T. Kortemme, *Design of multi-specificity in protein interfaces.* PLoS Comput Biol, 2007 Aug. **3**(8): p. e164.
2. Humphris, E.L. and T. Kortemme, *Prediction of protein-protein interface sequence diversity using flexible backbone computational protein design.* Structure, 2008 Dec 10. **16**(12): p. 1777--1788.
3. Desjarlais, J.R. and T.M. Handel, *De novo design of the hydrophobic cores of proteins.* Protein Sci, 1995. **4**(10): p. 2006-18.
4. Bolon, D.N., et al., *Prudent modeling of core polar residues in computational protein design.* J Mol Biol, 2003. **329**(3): p. 611-22.
5. Dahiyat, B.I. and S.L. Mayo, *De novo protein design: fully automated sequence selection.* Science, 1997. **278**(5335): p. 82-7.
6. Harbury, P.B., et al., *High-resolution protein design with backbone freedom.* Science, 1998. **282**(5393): p. 1462-7.
7. Kuhlman, B., et al., *Design of a novel globular protein fold with atomic-level accuracy.* Science, 2003. **302**(5649): p. 1364-8.
8. Shifman, J.M. and S.L. Mayo, *Modulating calmodulin binding specificity through computational protein design.* J Mol Biol, 2002. **323**(3): p. 417-23.
9. Chevalier, B.S., et al., *Design, activity, and structure of a highly specific artificial endonuclease.* Mol Cell, 2002. **10**(4): p. 895-905.

10. Havranek, J.J. and P.B. Harbury, *Automated design of specificity in molecular recognition.* Nat Struct Biol, 2003. **10**(1): p. 45-52.

11. Kortemme, T., et al., *Computational redesign of protein-protein interaction specificity.* Nat Struct Mol Biol, 2004. **11**(4): p. 371-9.

12. Kortemme, T. and D. Baker, *Computational design of protein-protein interactions.* Curr Opin Chem Biol, 2004. **8**(1): p. 91--97.

13. Bolon, D.N. and S.L. Mayo, *Enzyme-like proteins by computational design.* Proc Natl Acad Sci U S A, 2001. **98**(25): p. 14274-9.

14. Dwyer, M.A., L.L. Looger, and H.W. Hellinga, *Computational design of a biologically active enzyme.* Science, 2004. **304**(5679): p. 1967-71.

15. Havranek, J.J. and P.B. Harbury, *Automated design of specificity in molecular recognition.* Nat Struct Biol, 2003 Jan. **10**(1): p. 45--52.

16. Bolon, D.N., et al., *Specificity versus stability in computational protein design.* Proc Natl Acad Sci U S A, 2005. **102**(36): p. 12724-9.

17. Baldwin, E.P., et al., *The role of backbone flexibility in the accommodation of variants that repack the core of T4 lysozyme.* Science, 1993. **262**(5140): p. 1715-8.

18. Bordner, A.J. and R.A. Abagyan, *Large-scale prediction of protein geometry and stability changes for arbitrary single point mutations.* Proteins, 2004. **57**(2): p. 400-13.

19. Su, A. and S.L. Mayo, *Coupling backbone flexibility and amino acid sequence selection in protein design.* Protein Sci, 1997. **6**(8): p. 1701-7.

20. Harbury, P.B., et al., *High-resolution protein design with backbone freedom.* Science, 1998 Nov 20. **282**(5393): p. 1462--1467.

21. Desjarlais, J.R. and T.M. Handel, *Side-chain and backbone flexibility in protein core design.* J Mol Biol, 1999. **290**(1): p. 305-18.

22. Simons, K.T., et al., *Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions.* J Mol Biol, 1997. **268**(1): p. 209-25.

23. Simons, K.T., et al., *Ab initio protein structure prediction of CASP III targets using ROSETTA.* Proteins, 1999. **Suppl 3**: p. 171-6.

24. Kuhlman, B., et al., *Design of a novel globular protein fold with atomic-level accuracy.* Science, 2003 Nov 21. **302**(5649): p. 1364--1368.

25. Saunders, C.T. and D. Baker, *Recapitulation of protein family divergence using flexible backbone protein design.* J Mol Biol, 2005. **346**(2): p. 631-44.

26. Wei, Y. and M.H. Hecht, *Enzyme-like proteins from an unselected library of designed amino acid sequences.* Protein Eng Des Sel, 2004 Jan. **17**(1): p. 67--75.

27. Treynor, T.P., et al., *Computationally designed libraries of fluorescent proteins evaluated by preservation and diversity of function.* Proc Natl Acad Sci U S A, 2007. **104**(1): p. 48--53.

28. Kortemme, T. and D. Baker, *A simple physical model for binding energy hot spots in protein-protein complexes.* Proc Natl Acad Sci U S A, 2002. **99**(22): p. 14116-21.

29. Kortemme, T., A.V. Morozov, and D. Baker, *An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes.* J Mol Biol, 2003. **326**(4): p. 1239-59.

30. Lazaridis, T. and M. Karplus, *Effective energy function for proteins in solution.* Proteins, 1999. **35**(2): p. 133-52.

31. Davis, I.W., et al., *The backrub motion: how protein backbone shrugs when a sidechain dances.* Structure, 2006 Feb. **14**(2): p. 265--274.

32. Davis, I.W., et al., *The backrub motion: how protein backbone shrugs when a sidechain dances.* Structure, 2006. **14**(2): p. 265-74.

33. Smith, C.A. and T. Kortemme, *Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction.* J Mol Biol, 2008. **380**(4): p. 742-56.

34. DePristo, M.A., D.M. Weinreich, and D.L. Hartl, *Missense meanderings in sequence space: a biophysical view of protein evolution.* Nat Rev Genet, 2005. **6**(9): p. 678-87.

35. Kim, P.M., et al., *Relating three-dimensional structures to protein networks provides evolutionary insights.* Science, 2006. **314**(5807): p. 1938-41.

36. Kortemme, T. and D. Baker, *Computational design of protein-protein interactions.* Curr Opin Chem Biol, 2004. **8**(1): p. 91-7.

37. Watters, A.L., et al., *The highly cooperative folding of small naturally occurring proteins is likely the result of natural selection.* Cell, 2007. **128**(3): p. 613-24.

38. Zarrinpar, A., S.H. Park, and W.A. Lim, *Optimization of specificity in a cellular protein interaction network by negative selection.* Nature, 2003. **426**(6967): p. 676-80.

39. Ambroggio, X.I. and B. Kuhlman, *Computational design of a single amino acid sequence that can switch between two distinct protein folds.* J Am Chem Soc, 2006. **128**(4): p. 1154-61.

40. Bogan, A.A. and K.S. Thorn, *Anatomy of hot spots in protein interfaces.* J Mol Biol, 1998. **280**(1): p. 1-9.

41. Clackson, T. and J.A. Wells, *A hot spot of binding energy in a hormone-receptor interface.* Science, 1995. **267**(5196): p. 383-6.

42. Lowman, H.B. and J.A. Wells, *Affinity maturation of human growth hormone by monovalent phage display.* J Mol Biol, 1993. **234**(3): p. 564-78.

43. Pal, G., et al., *Comprehensive and quantitative mapping of energy landscapes for protein-protein interactions by rapid combinatorial scanning.* J Biol Chem, 2006. **281**(31): p. 22378-85.

44. Andreeva, A., et al., *SCOP database in 2004: refinements integrate structure and sequence family data.* Nucleic Acids Res, 2004. **32**(Database issue): p. D226-9.

45. Xenarios, I., et al., *DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions.* Nucleic Acids Res, 2002. **30**(1): p. 303-5.

46. Davis, F.P. and A. Sali, *PIBASE: a comprehensive database of structurally defined protein interfaces.* Bioinformatics, 2005. **21**(9): p. 1901-7.

47. Glaser, F., et al., *The ConSurf-HSSP database: the mapping of evolutionary conservation among homologs onto PDB structures.* Proteins, 2005. **58**(3): p. 610-7.

48. DeLano, W.L., et al., *Convergent solutions to binding at a protein-protein interface.* Science, 2000. **287**(5456): p. 1279-83.

49. Keskin, O. and R. Nussinov, *Similar binding sites and different partners: implications to shared proteins in cellular pathways.* Structure, 2007. **15**(3): p. 341-54.

50. Jackson, R.M., *Comparison of protein-protein interactions in serine protease-inhibitor and antibody-antigen complexes: implications for the protein docking problem.* Protein Sci, 1999. **8**(3): p. 603-13.

51. Lu, W., et al., *Binding of amino acid side-chains to S1 cavities of serine proteinases.* J Mol Biol, 1997. **266**(2): p. 441-61.

52. Lo Conte, L., C. Chothia, and J. Janin, *The atomic structure of protein-protein recognition sites.* J Mol Biol, 1999. **285**(5): p. 2177-98.

53. Ma, B., et al., *Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces.* Proc Natl Acad Sci U S A, 2003. **100**(10): p. 5772-7.

54. Villar, H.O. and L.M. Kauvar, *Amino acid preferences at protein binding sites.* FEBS Lett, 1994. **349**(1): p. 125-30.

55. Mossessova, E., R.A. Corpina, and J. Goldberg, *Crystal structure of ARF1*Sec7 complexed with Brefeldin A and its implications for the guanine nucleotide exchange mechanism.* Mol Cell, 2003. **12**(6): p. 1403-11.

56. Koehl, P. and M. Levitt, *Protein topology and stability define the space of allowed sequences.* Proc Natl Acad Sci U S A, 2002. **99**(3): p. 1280-5.

57. Taverna, D.M. and R.A. Goldstein, *Why are proteins so robust to site mutations?* J Mol Biol, 2002. **315**(3): p. 479-84.

58. Wagner, A., *Robustness, evolvability, and neutrality.* FEBS Lett, 2005. **579**(8): p. 1772-8.

59. Aharoni, A., et al., *The 'evolvability' of promiscuous protein functions.* Nat Genet, 2005. **37**(1): p. 73-6.

60. Pokala, N. and T.M. Handel, *Review: protein design--where we were, where we are, where we're going.* J Struct Biol, 2001. **134**(2-3): p. 269-81.

61. Xia, Y. and M. Levitt, *Simulating protein evolution in sequence and structure space.* Curr Opin Struct Biol, 2004. **14**(2): p. 202-7.

62. Butterfoss, G.L. and B. Kuhlman, *Computer-based design of novel protein structures.* Annu Rev Biophys Biomol Struct, 2006. **35**: p. 49-65.

63. Ponder, J.W. and F.M. Richards, *Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes.* J Mol Biol, 1987. **193**(4): p. 775-91.

64. Larson, S.M., et al., *Thoroughly sampling sequence space: large-scale protein design of structural ensembles.* Protein Sci, 2002. **11**(12): p. 2804-13.

65. Dokholyan, N.V. and E.I. Shakhnovich, *Understanding hierarchical protein evolution from first principles.* J Mol Biol, 2001. **312**(1): p. 289-307.

66. Jaramillo, A., et al., *Folding free energy function selects native-like protein sequences in the core but not on the surface.* Proc Natl Acad Sci U S A, 2002. **99**(21): p. 13554-9.

67. Koehl, P. and M. Levitt, *De novo protein design. II. Plasticity in sequence space.* J Mol Biol, 1999. **293**(5): p. 1183-93.

68. Kuhlman, B. and D. Baker, *Native protein sequences are close to optimal for their structures.* Proc Natl Acad Sci U S A, 2000. **97**(19): p. 10383-8.

69. Hayes, R.J., et al., *Combining computational and experimental screening for rapid optimization of protein properties.* Proc Natl Acad Sci U S A, 2002. **99**(25): p. 15926-31.

70. Treynor, T.P., et al., *Computationally designed libraries of fluorescent proteins evaluated by preservation and diversity of function.* Proc Natl Acad Sci U S A, 2007. **104**(1): p. 48-53.

71. Fu, X., J.R. Apgar, and A.E. Keating, *Modeling Backbone Flexibility to Achieve Sequence Diversity: The Design of Novel alpha-Helical Ligands for Bcl-x(L).* J Mol Biol, 2007.

72. Ding, F. and N.V. Dokholyan, *Emergence of protein fold families through rational design.* PLoS Comput Biol, 2006. **2**(7): p. e85.

73. Humphris, E.L. and T. Kortemme, *Design of multi-specificity in protein interfaces.* PLoS Comput Biol, 2007. **3**(8): p. e164.

74. Dunbrack, R.L., Jr., *Rotamer libraries in the 21st century.* Curr Opin Struct Biol, 2002. **12**(4): p. 431-40.

75. Cunningham, B.C. and J.A. Wells, *Rational design of receptor-specific variants of human growth hormone.* Proc Natl Acad Sci U S A, 1991. **88**(8): p. 3407-11.

76. Sidhu, S.S., et al., *Phage display for selection of novel binding peptides.* Methods Enzymol, 2000. **328**: p. 333-63.

77. Schiffer, C., et al., *Structure of a phage display-derived variant of human growth hormone complexed to two copies of the extracellular domain of its receptor: evidence for strong structural coupling between receptor binding sites.* J Mol Biol, 2002. **316**(2): p. 277-89.

78. Friedland, G.D., et al., *A correspondence between solution-state dynamics of an individual protein and the sequence and conformational diversity of its family.* PLoS Comput Biol, 2009. **5**(5): p. e1000393.

79. Friedland, G.D., et al., *A simple model of backbone flexibility improves modeling of side-chain conformational variability.* J Mol Biol, 2008. **380**(4): p. 757-74.

80. Pal, G., et al., *Intramolecular cooperativity in a protein binding site assessed by combinatorial shotgun scanning mutagenesis.* J Mol Biol, 2005. **347**(3): p. 489-94.

81. Wollacott, A.M. and J.R. Desjarlais, *Virtual interaction profiles of proteins.* J Mol Biol, 2001. **313**(2): p. 317-42.

82. Eyre-Walker, A. and P.D. Keightley, *The distribution of fitness effects of new mutations.* Nat Rev Genet, 2007. **8**(8): p. 610--618.

83. James, L.C. and D.S. Tawfik, *Conformational diversity and protein evolution--a 60-year-old hypothesis revisited.* Trends Biochem Sci, 2003. **28**(7): p. 361-8.

84. Copley, S.D., *Enzymes with extra talents: moonlighting functions and catalytic promiscuity.* Curr Opin Chem Biol, 2003. **7**(2): p. 265-72.

85. Gupta, R.D. and D.S. Tawfik, *Directed enzyme evolution via small and effective neutral drift libraries.* Nat Methods, 2008. **5**(11): p. 939-42.

86. Zhao, H., *Directed evolution of novel protein functions.* Biotechnol Bioeng, 2007 Oct 1. **98**(2): p. 313--317.

87. Aharoni, A., et al., *The 'evolvability' of promiscuous protein functions.* Nat Genet, 2005. **37**(1): p. 73-6.

88. Fernández, A., et al., *Protein promiscuity: drug resistance and native functions--HIV-1 case.* J Biomol Struct Dyn, 2005. **22**(6): p. 615-24.

89. Rhee, S.-Y., et al., *Human immunodeficiency virus reverse transcriptase and protease sequence database.* Nucleic Acids Res, 2003. **31**(1): p. 298-303.

90. Carrillo, A., et al., *In vitro selection and characterization of human immunodeficiency virus type 1 variants with increased resistance to ABT-378, a novel protease inhibitor.* J Virol, 1998. **72**(9): p. 7532-41.

91. Seelmeier, S., et al., *Human immunodeficiency virus has an aspartic-type protease that can be inhibited by pepstatin A.* Proc Natl Acad Sci U S A, 1988. **85**(18): p. 6612-6.

92. Hill, M., G. Tachedjian, and J. Mak, *The packaging and maturation of the HIV-1 Pol proteins.* Curr HIV Res, 2005 Jan. **3**(1): p. 73--85.

93. Debouck, C., et al., *Human immunodeficiency virus protease expressed in Escherichia coli exhibits autoprocessing and specific maturation of the gag precursor.* Proc Natl Acad Sci U S A, 1987. **84**(24): p. 8903-6.

94. Ho, D.D., et al., *Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection.* Nature, 1995. **373**(6510): p. 123-6.

95. Wei, X., et al., *Viral dynamics in human immunodeficiency virus type 1 infection.* Nature, 1995. **373**(6510): p. 117-22.

96. Coffin, J.M., *HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy.* Science, 1995. **267**(5197): p. 483-9.

97. Wu, T.D., et al., *Mutation patterns and structural correlates in human immunodeficiency virus type 1 protease following different protease inhibitor treatments.* J Virol, 2003 Apr. **77**(8): p. 4836--4847.

98. Condra, J.H., et al., *In vivo emergence of HIV-1 variants resistant to multiple protease inhibitors.* Nature, 1995 Apr 6. **374**(6522): p. 569--571.

99. Molla, A., et al., *Ordered accumulation of mutations in HIV protease confers resistance to ritonavir.* Nat Med, 1996 Jul. **2**(7): p. 760--766.

100. Ishikita, H. and A. Warshel, *Predicting drug-resistant mutations of HIV protease.* Angew Chem Int Ed Engl, 2008. **47**(4): p. 697-700.

101. Lin, Y., et al., *Effect of point mutations on the kinetics and the inhibition of human immunodeficiency virus type 1 protease: relationship to drug resistance.* Biochemistry, 1995 Jan 31. **34**(4): p. 1143--1152.

102. Muzammil, S., P. Ross, and E. Freire, *A major role for a set of non-active site mutations in the development of HIV-1 protease drug resistance.* Biochemistry, 2003. **42**(3): p. 631--638.

103. Olsen, D.B., et al., *Non-active site changes elicit broad-based cross-resistance of the HIV-1 protease to inhibitors.* J Biol Chem, 1999. **274**(34): p. 23699-701.

104. Piana, S., P. Carloni, and U. Rothlisberger, *Drug resistance in HIV-1 protease: Flexibility-assisted mechanism of compensatory mutations.* Protein Sci, 2002 Oct. **11**(10): p. 2393--2402.

105. Baldwin, E.T., et al., *Structural basis of drug resistance for the V82A mutant of HIV-1 proteinase.* Nat Struct Biol, 1995 Mar. **2**(3): p. 244--249.

106. Loeb, D.D., et al., *Complete mutagenesis of the HIV-1 protease.* Nature, 1989. **340**(6232): p. 397-400.

107. Sadiq, S.K., et al., *Automated molecular simulation based binding affinity calculator for ligand-bound HIV-1 proteases.* J Chem Inf Model, 2008 Sep. **48**(9): p. 1909--1919.

108. Rhee, S.-Y., et al., *HIV-1 Protease and reverse-transcriptase mutations: correlations with antiretroviral therapy in subtype B isolates and implications for drug-resistance surveillance.* J Infect Dis, 2005. **192**(3): p. 456-65.

109. Bennett, D.E., et al., *Drug resistance mutations for surveillance of transmitted HIV-1 drug-resistance: 2009 update.* PLoS One, 2009. **4**(3): p. e4724.

110. Johnson, V.A., et al., *Update of the Drug Resistance Mutations in HIV-1.* Top HIV Med, 2008. **16**(5): p. 138-45.

111. Altman, M.D., et al., *HIV-1 protease inhibitors from inverse design in the substrate envelope exhibit subnanomolar binding to drug-resistant variants.* J Am Chem Soc, 2008 May 14. **130**(19): p. 6099--6113.

112. Chellappan, S., et al., *Evaluation of the substrate envelope hypothesis for inhibitors of HIV-1 protease.* Proteins, 2007 Aug 1. **68**(2): p. 561--567.

113. Zennou, V., et al., *Loss of viral fitness associated with multiple Gag and Gag-Pol processing defects in human immunodeficiency virus type 1 variants selected for resistance to protease inhibitors in vivo.* J Virol, 1998 Apr. **72**(4): p. 3300--3306.

114. Prabu-Jeyabalan, M., et al., *Structural basis for coevolution of a human immunodeficiency virus type 1 nucleocapsid-p1 cleavage site with a V82A drug-resistant mutation in viral protease.* J Virol, 2004 Nov. **78**(22): p. 12446--12454.

115. Zhang, Y.M., et al., *Drug resistance during indinavir therapy is caused by mutations in the protease gene and in its Gag substrate cleavage sites.* J Virol, 1997. **71**(9): p. 6662-70.

116. Dauber, D.S., et al., *Altered substrate specificity of drug-resistant human immunodeficiency virus type 1 protease.* J Virol, 2002 Feb. **76**(3): p. 1359--1368.

117. Cote, H.C., Z.L. Brumme, and P.R. Harrigan, *Human immunodeficiency virus type 1 protease cleavage site mutations associated with protease inhibitor cross-resistance selected by indinavir, ritonavir, and/or saquinavir.* J Virol, 2001 Jan. **75**(2): p. 589--594.

118. Bally, F., et al., *Polymorphism of HIV type 1 gag p7/p1 and p1/p6 cleavage sites: clinical significance and implications for resistance to protease inhibitors.* AIDS Res Hum Retroviruses, 2000. **16**(13): p. 1209-13.

119. Nijhuis, M., et al., *Increased fitness of drug resistant HIV-1 protease as a result of acquisition of compensatory mutations during suboptimal therapy.* AIDS, 1999. **13**(17): p. 2349--2359.

120. Liu, Y., E. Eyal, and I. Bahar, *Analysis of correlated mutations in HIV-1 protease using spectral clustering.* Bioinformatics, 2008. **24**(10): p. 1243-50.

121. Garriga, C., et al., *Mutational patterns and correlated amino acid substitutions in the HIV-1 protease after virological failure to nelfinavir- and lopinavir/ritonavir-based treatments.* J Med Virol, 2007. **79**(11): p. 1617--1628.

122. Hong, L., et al., *Crystal structure of an in vivo HIV-1 protease mutant in complex with saquinavir: insights into the mechanisms of drug resistance.* Protein Sci, 2000. **9**(10): p. 1898--1904.

123. Jenwitheesuk, E. and R. Samudrala, *Improved prediction of HIV-1 protease-inhibitor binding energies by molecular dynamics simulations.* BMC Struct Biol, 2003. **3**: p. 2.

124. Jenwitheesuk, E. and R. Samudrala, *Prediction of HIV-1 protease inhibitor resistance using a protein-inhibitor flexible docking approach.* Antivir Ther, 2005. **10**(1): p. 157--166.

125. Best, R.B., et al., *Relation between native ensembles and experimental structures of proteins.* Proc Natl Acad Sci U S A, 2006. **103**(29): p. 10901-6.

126. Hornak, V., et al., *HIV-1 protease flaps spontaneously open and reclose in molecular dynamics simulations.* Proc Natl Acad Sci U S A, 2006 Jan 24. **103**(4): p. 915--920.

127. Erickson, J.W. and S.K. Burt, *Structural mechanisms of HIV drug resistance.* Annu Rev Pharmacol Toxicol, 1996. **36**: p. 545-71.

128. Rose, R.B., C.S. Craik, and R.M. Stroud, *Domain flexibility in retroviral proteases: structural implications for drug resistant mutations.* Biochemistry, 1998 Feb 24. **37**(8): p. 2607--2621.

129. Ho, D.D., et al., *Characterization of human immunodeficiency virus type 1 variants with increased resistance to a C2-symmetric protease inhibitor.* J Virol, 1994 Mar. **68**(3): p. 2016--2020.

130. Kaplan, A.H., et al., *Selection of multiple human immunodeficiency virus type 1 variants that encode viral proteases with decreased sensitivity to an inhibitor of the viral protease.* Proc Natl Acad Sci U S A, 1994 Jun 7. **91**(12): p. 5597--5601.

131. Perryman, A.L., J.-H. Lin, and J.A. McCammon, *HIV-1 protease molecular dynamics of a wild-type and of the V82F/I84V mutant: possible contributions to drug resistance and a potential new target site for drugs.* Protein Sci, 2004 Apr. **13**(4): p. 1108--1123.

132. Wang, W. and P.A. Kollman, *Computational study of protein specificity: the molecular basis of HIV-1 protease drug resistance.* Proc Natl Acad Sci U S A, 2001. **98**(26): p. 14937-42.

133. Chou, K.C., et al., *Predicting human immunodeficiency virus protease cleavage sites in proteins by a discriminant function method.* Proteins, 1996. **24**(1): p. 51-72.

134. Chou, K.C., C.T. Zhang, and F.J. Kezdy, *A vector projection approach to predicting HIV protease cleavage sites in proteins.* Proteins, 1993. **16**(2): p. 195-204.

135. Chellappan, S., et al., *Design of mutation-resistant HIV protease inhibitors with the substrate envelope hypothesis.* Chem Biol Drug Des, 2007. **69**(5): p. 298-313.

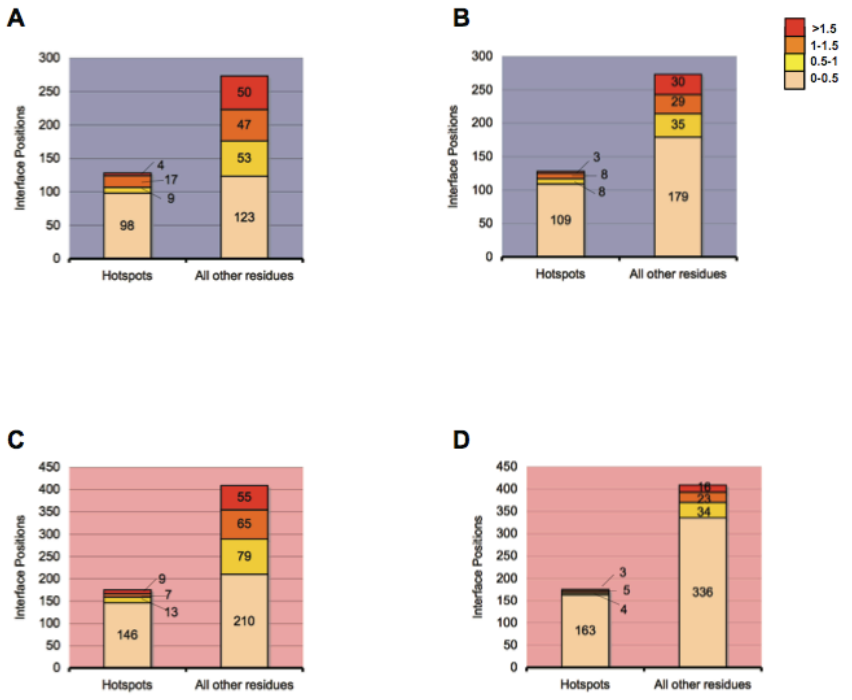
136. Ohtaka, H. and E. Freire, *Adaptive inhibitors of the HIV-1 protease.* Prog Biophys Mol Biol, 2005. **88**(2): p. 193-208.

137. Clemente, J.C., et al., *Design, synthesis, evaluation, and crystallographic-based structural studies of HIV-1 protease inhibitors with reduced response to the V82A mutation.* Journal of Medicinal Chemistry, 2008. **51**(4): p. 852-860.

138. Babor, M. and T. Kortemme, *Multi-constraint computational design suggests that native sequences of germline antibody H3 loops are nearly optimal for conformational flexibility.* Proteins, 2009. **75**(4): p. 846-58.

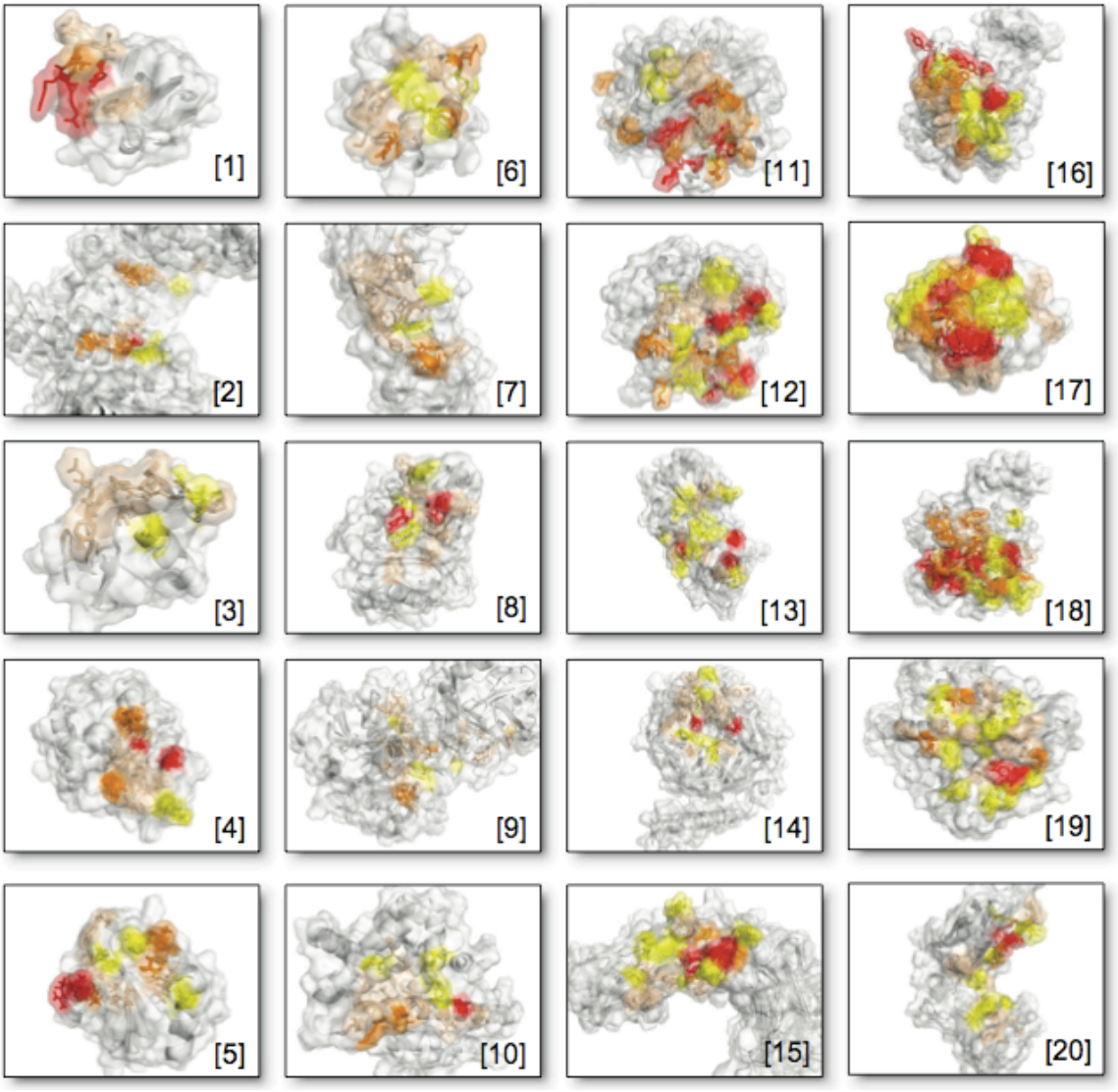
Appendix A. Chapter 2 Supplementary Materials

A.1.1. Supplementary Figures

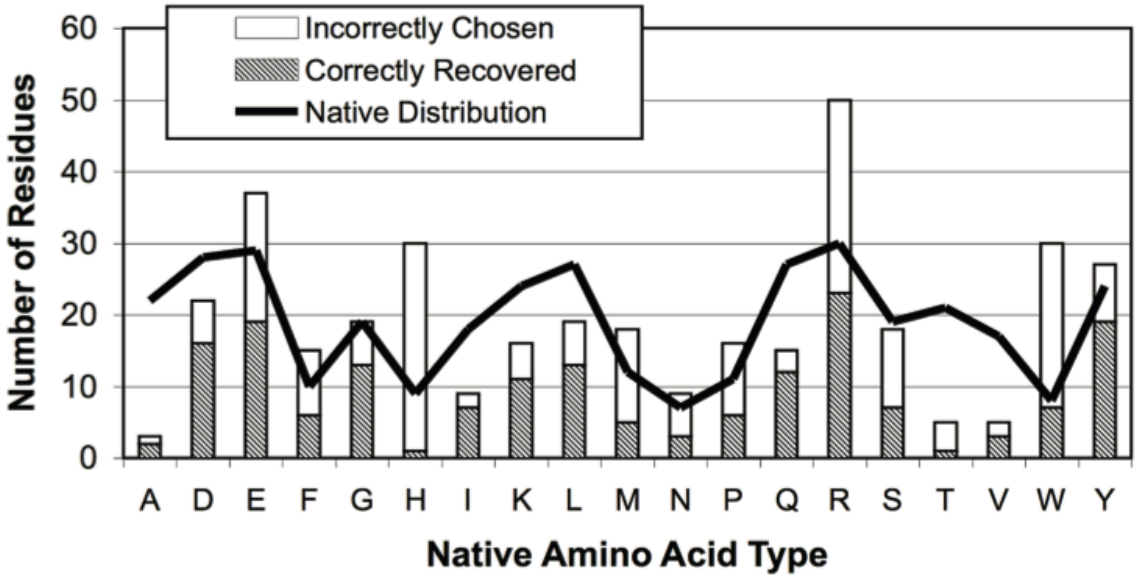


Supplementary Figure A.1 Group I and Group II distributions of optimization in promiscuous interfaces.
Predicted per-residue binding score improvements (relative to native) are shown for sequences selected in single- **(A,C)** and multi- **(B,D)** constraint simulations for group I (top, pink shading) and group II (bottom, blue shading). Colored bars indicate the magnitude of predicted per-residue improvement over native. Darker colored bars (compromise value >1, orange, red) indicate positions for which the simulation predicts a non-native residue to bind stronger than native. Lighter colored bars (compromise value <1, wheat, yellow) indicate simulations recovered the native (or near-native) residue type. Group I and group II

proteins show similar distributions of native residues predicted to be suboptimal when optimized for single binding interactions alone (**A**, **C**; compare red, orange, yellow bars). Optimization over multiple partners, however, differed between groups: a larger number of non-hotspot positions were still predicted to be suboptimal for group I when all partners were considered for optimization than seen in group II ("all other residues", compare red, orange, yellow bars in **B**, **D**). This is consistent with our finding that native sequence recovery is lower overall for group I.



Supplementary Figure A.2 Distribution of compromise for all 20 promiscuous proteins in the dataset. Constraint scores (see Section 2.4.4) are mapped onto each promiscuous protein in the dataset. Darker colors indicate stronger tradeoff in that some partners considered are predicted to "give up" potential gain so that other partners could fulfill their optimal interactions Overall, group I proteins (1-10,20) display lower levels of tradeoff than seen in group II (11-19).



Supplementary Figure A.3 Amino acid frequency distributions of sequences selected as optimal in the multi-constraint procedure.
For each amino acid type, the number of times an amino acid type was correctly recovered as native is shown as black striped bars. Non-native substitutions of each amino acid type are shown as white bars. The native amino acid distribution is plotted for reference (solid black line).

1 Design Protocol — FYN SH3 Interface Residues

	93	96	97	98	99	100	119	Identical to Native #	%
NATIVE	Y	R	T	E	D	D	W	7	
MULTI	R	+	P	W	+	+	+	4	57%
SINGLE 1AVZ	R	+	S	Q	E	+	+	3	43%
1M27	R	E	+	W	+	E	+	3	43%

2 Design Protocol — Importin Beta Interface Residues

	55	106	156	159	239	274	278	281	288	342	Identical to Native #	%
NATIVE	V	P	Y	Q	Q	E	Q	E	D	W	10	
MULTI	L	K	+	E	+	W	+	+	Y	M	4	40%
SINGLE 1IBR	L	M	+	E	R	H	+	+	Y	+	4	40%
1M5N	E	W	+	K	I	Y	W	Y	W	E	1	10%

3 Design Protocol — Ovomucoid Inhibitor Interface Residues

	13	14	15	17	18	19	20	21	33	36	Identical to Native #	%
NATIVE	K	P	A	T	K	E	Y	R	N	N	10	
MULTI	W	W	P	E	F	H	W	+	I	V	1	10%
SINGLE 1HJA	W	W	P	E	F	R	F	+	I	S	1	10%
1SGP	W	W	R	E	Y	+	R	Q	E	R	1	10%

4 Design Protocol — CheY Interface Residues

	90	92	95	99	104	105	106	107	108	119	122	Identical to Native #	%
NATIVE	A	K	I	A	S	G	Y	V	V	K	K	11	
MULTI	W	+	+	G	P	S	+	F	S	+	R	4	36%
SINGLE 1F4V	R	R	+	W	N	S	H	W	S	+	I	2	18%
1FFG	H	+	Q	+	P	H	+	W	F	W	+	4	36%
1KMI	T	W	+	G	T	N	H	H	W	+	R	2	18%

5 Design Protocol — Phosphocarrier Protein HPR Interface Residues

	11	12	15	16	47	48	51	52	53	54	55	Identical to Native #	%
NATIVE	D	S	H	A	I	M	M	S	L	G	I	11	
MULTI	R	R	+	W	H	+	F	+	H	+	F	4	36%
SINGLE 1KKL	R	G	R	S	R	+	W	+	H	+	Y	3	27%
1RZR	H	W	+	W	H	L	F	Q	W	+	E	2	18%

6 Design Protocol — Thioredoxin Interface Residues

	31	34	37	60	72	73	74	75	91	93	94	98	Identical to Native #	%
NATIVE	W	P	M	I	I	R	G	I	V	A	L	Q	12	
MULTI	+	+	R	H	E	W	+	+	W	L	Q	W	4	33%
SINGLE 1F6M	+	M	R	+	M	W	+	N	W	Q	W	R	3	25%
1T7P	+	+	W	F	W	+	H	T	+	I	Q	W	4	33%

7 Design Protocol — IL-6 Interface Residues

	142	144	145	147	165	166	167	168	169	170	193	226	229	230	Identical to Native #	%
NATIVE	W	T	H	F	S	T	V	Y	F	V	D	S	S	V	14	
MULTI	+	R	R	+	+	H	I	+	+	F	F	K	H	+	6	43%
SINGLE 1I1R	+	R	R	+	+	E	I	+	+	F	R	K	H	D	5	36%
1P9M	+	E	R	R	A	H	I	P	+	+	F	E	S	F	3	21%

8 Design Protocol — Beta Lactamase Interface Residues

	99	100	102	103	104	105	106	107	110	111	129	216	237	270	Identical to Native #	%
NATIVE	Q	N	L	V	E	Y	S	P	E	K	M/L	V	A	M	14	
MULTI	H	W	R	E	+	+	G	H	+	+	Q	S	G	W	4	29%
SINGLE 1JTD	+	W	H	H	P	W	L	H	+	+	Q	H	G	W	3	21%
1JTG	F	W	W	E	+	+	K	H	H	Q	Q	S	N	W	2	14%

9 Design Protocol — Elastase Interface Residues

	41	57	61	172	175	190	192	193	194	195	213	214	215	216	217	Identical to Native #	%
NATIVE	T	H	R	W	T	G	Q	G	D	S	T	S	F	V	S	15	
MULTI	L	F	+	+	K	+	+	D	P	L	A	Y	Y	T	N	4	27%
SINGLE 1EAI	L	F	V	+	K	+	+	N	P	I	M	K	W	T	W	3	20%
1FLE	L	F	M	Y	W	Y	+	N	V	L	L	Y	+	I	W	2	13%
1MCV	+	F	+	+	W	+	+	H	P	L	K	L	Y	A	P	5	33%

10 Design Protocol — Peroxisome Proliferator Receptor Interface Residues

	396	410	414	430	432	433	434	436	437	439	440	443	444	447	451	Identical to Native #	%
NATIVE	D	Q	L	Q	F	A	K	L	Q	M	T	R	Q	T	Q	15	
MULTI	W	+	+	+	Y	+	R	H	E	K	E	P	W	H	+	5	33%
SINGLE 1K74	W	M	+	K	R	P	W	Y	E	H	Y	D	W	W	N	1	7%
2PRG	+	+	H	+	Y	+	W	H	W	K	D	F	E	Q	+	5	33%

11 Design Protocol — Ran Interface Residues

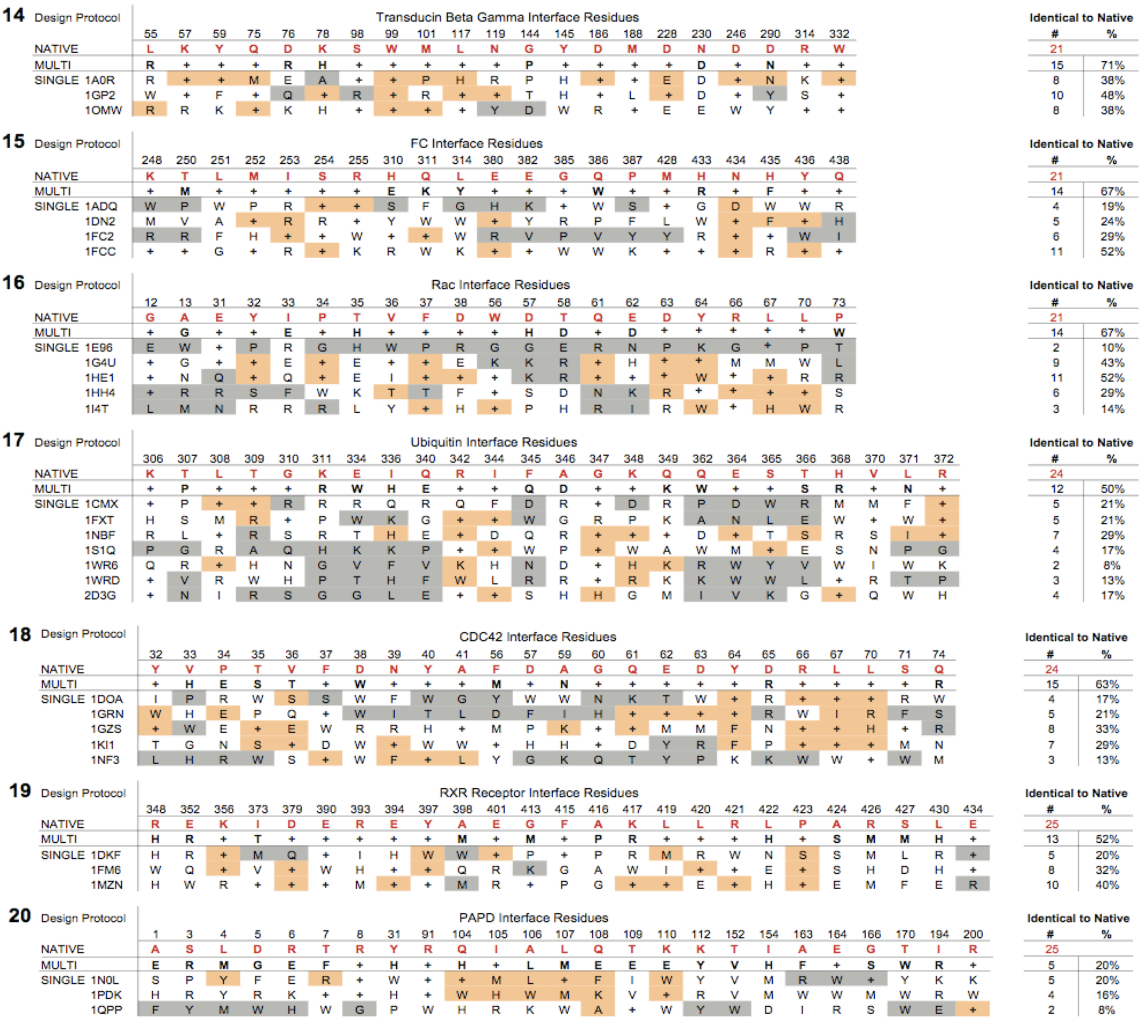
	43	44	45	67	68	70	71	74	75	76	77	78	79	95	96	106	110	134	Identical to Native #	%
NATIVE	L	G	V	A	G	E	K	G	L	R	D	G	Y	R	V	R	R	K	18	
MULTI	E	+	M	G	+	H	+	+	+	+	+	+	A	+	H	+	+	+	12	67%
SINGLE 1A2K	Q	+	W	V	+	W	+	H	G	+	R	+	Q	L	R	Y	M	V	5	28%
1I2M	R	R	E	H	+	R	E	+	G	+	R	L	V	+	H	+	F	+	6	33%
1IBR	K	M	L	T	+	W	W	S	+	+	+	T	F	N	G	Q	+	W	5	28%
1K5D	+	+	P	G	F	Y	R	D	+	L	I	F	G	+	R	G	Y	R	4	22%
1WA5	E	S	E	T	Y	R	D	+	+	+	+	M	E	+	K	+	E	+	7	39%

12 Design Protocol — Ras Interface Residues

	13	21	25	31	32	33	34	36	37	38	39	40	41	54	61	63	64	67	70	Identical to Native #	%
NATIVE	G	I	Q	E	Y	D	P	I	E	D	S	Y	R	D	Q	E	Y	M	Q	19	
MULTI	D	L	R	R	H	+	+	+	+	+	+	+	+	+	+	+	+	H	+	13	68%
SINGLE 1BKD	H	+	F	W	H	E	S	V	+	+	N	+	W	+	+	+	+	H	+	9	47%
1HE8	L	M	R	R	W	E	K	R	+	P	+	H	W	I	G	W	H	W	K	2	11%
1K8R	P	R	W	M	S	W	V	E	+	+	+	D	+	+	T	K	+	W	K	6	32%
1LFD	+	F	R	R	R	+	N	W	+	E	R	W	M	W	E	W	T	M	R	3	16%
1WQ1	D	R	R	+	H	+	+	+	M	+	H	S	W	W	+	+	+	W	W	8	42%

13 Design Protocol — Actin Interface Residues

	113	143	144	145	146	147	148	166	167	169	173	286	287	288	290	345	355	371	372	373	Identical to Native #	%
NATIVE	K	Y	A	S	G	R	T	Y	E	Y	H	D	V	D	R	I	M	H	R	K	20	
MULTI	R	+	S	R	+	+	S	+	H	+	E	+	H	W	+	+	W	R	N	N	8	40%
SINGLE 1EQY	S	+	S	F	S	P	S	S	W	+	F	R	A	P	L	+	F	Q	E	W	3	15%
1HLU	R	W	I	H	F	T	+	+	W	+	F	+	R	+	W	H	W	R	+	M	6	30%
1LOT	R	F	L	R	+	+	M	+	H	Q	E	H	H	W	+	F	W	Q	N	H	4	20%



Supplementary Figure A.4 Single- and Multi- Constraint Sequences Selected for 20 Multi-Specific Proteins

The first row in each Table contains the interface residue PDB numbering, the second row lists the native sequence (red), and the following rows contain sequences predicted to be optimal in each simulation: multi-constraint (2nd sequence), single constraint (3rd - last sequences). Plus signs in the Table denote that the native amino acid residue type was recovered as optimal. The number and percent of interface residues recovered to be identical to native is shown for each simulation in the rightmost column. Pink shading denotes that the original wild type amino acid type was calculated to be a hotspot for the given binding partner while grey shading signifies an interface positions not within 4Å of the shaded interaction partner (see Methods section in the main manuscript).

A.1.2. Supplementary Tables

	PROMISCUOUS PROTEIN	DIP ID NUMBER	ORGANISM	EVALUE	DIP PROTEIN DESCRIPTION
1	FYN SH3	DIP:198N	*Mus musculus*	2.00E-29	Protein-tyrosine Kinase Fyn
2	IMPORTIN BETA	DIP:2357N	*Saccharomyces cerevisiae*	3.00E-82	KAP95 Protein
3	OVOMUCOID INHIBITOR	N/A	N/A	N/A	N/A
4	CHEY	DIP:6052N	*Escherichia coli*	8.00E-38	Chemotaxis Protein CheY
5	THIOREDOXIN	DIP:5552N	*Saccharomyces cerevisiae*	4.00E-12	Thioredoxin I
6	HPR	DIP:6180N	*Escherichia coli*	1.00E-09	Phosphocarrier protein HPr (Histidine-containing protein)
7	IL6	DIP:95N	*Homo sapiens*	1.00E-179	Membrane Glycoprotein GP130 Precursor
8	BETA LACTAMASE	N/A	N/A	N/A	N/A
9	ELASTASE	DIP:378N	*Sus scrofa domestica*	1.00E-132	Pancreatic Elastase I Precursor
10	PPR	DIP:241N	*Homo sapiens*	6.00E-97	Peroxisome Proliferator-activated Receptor
11	RAN	DIP:1357N	*Saccharomyces cerevisiae*	1.00E-101	GTP-binding protein GSP2
12	RAS	DIP:2263N	*Saccharomyces cerevisiae*	3.00E-47	GTP-binding protein RAS2
13	ACTIN	DIP:310N	*Saccharomyces cerevisiae*	0	actin
14	TRANSDUCIN BETA GAMMA	DIP:954N	*Saccharomyces cerevisiae*	2.00E-66	GTP-binding protein beta chain STE4
15	FC IGG1	N/A	N/A	N/A	N/A
16	RAC	DIP:862N	*Drosophila melanogaster*	1.00E-96	GTP-binding Protein Rac2
17	UBIQUITIN	DIP:1549N	*Saccharomyces cerevisiae*	9.00E-32	Ubiquitin
18	CDC42	DIP:862N	*Saccharomyces cerevisiae*	4.00E-87	Cell Division Control Protein CDC42
19	RXR RECEPTOR	DIP:641N	*Homo sapiens*	1.00E-138	Retinoic Acid Receptor X-alpha
20	PAPD	DIP:6196N	*Escherichia coli*	1.00E-120	Chaperone Protein PapD Precursor

Supplementary Table A.1 Source of high-throughput interaction data for promiscuous proteins. Database of Interacting Proteins (DIP http://dip.doe-mbi.ucla.edu/) identification numbers, e-values and protein names for sequences identified as homologues to the 20 promiscuous proteins in our dataset. Interaction graphs (see Figure 2.2) are taken directly from each DIP protein listed.

Appendix B. Chapter 3, Supplementary Materials

B.1.1. Supplementary Text

Filtering of Backbones and Variation in Thresholds

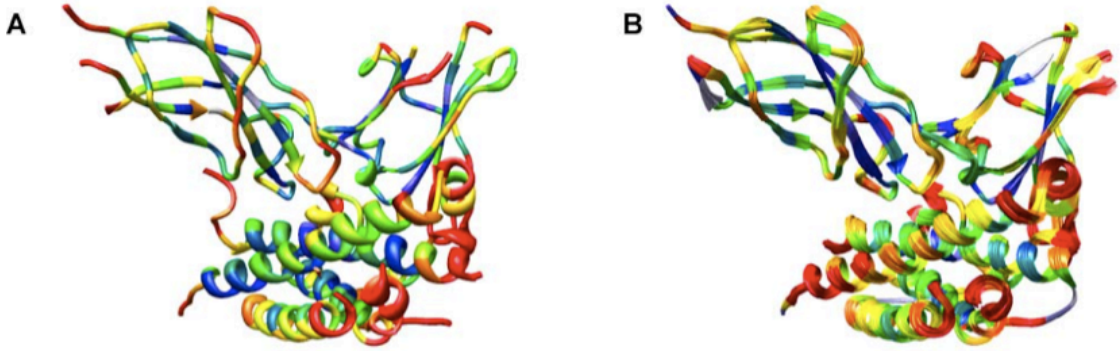
In order to estimate the influence of tunable parameters on the predicted tolerance profiles, we selected and scored sequences used a variety of conditions including: (i) design on up 500 "near-native" backbones, (ii) varying binding and folding thresholds to 0%, 5% and 10% of the score of the wild-type sequence, (iii) use of a weakened Leonard Jones repulsive term ("soft" repulsive), (iv) pre-filtering the ensemble of near-native backbones and designing only on those with most favorable (negative) scores and (v) running the genetic algorithm optimization for varying numbers of generations. Overall varying the parameters listed above resulted in qualitatively similar profiles. The one exception to this general robustness of the methodology to parameters is the effect of varying the folding threshold: if this threshold was too lenient (accepting sequences with scores differing from native by 10% or more) profiles which should show strong signal for a small set of amino acid residues became flattened out. Similarly if this threshold

was too strong (0%) only a very small number of sequences were selected for inclusion in the tolerance profiles, which often resulted in biases in distributions due to taking a small sample.

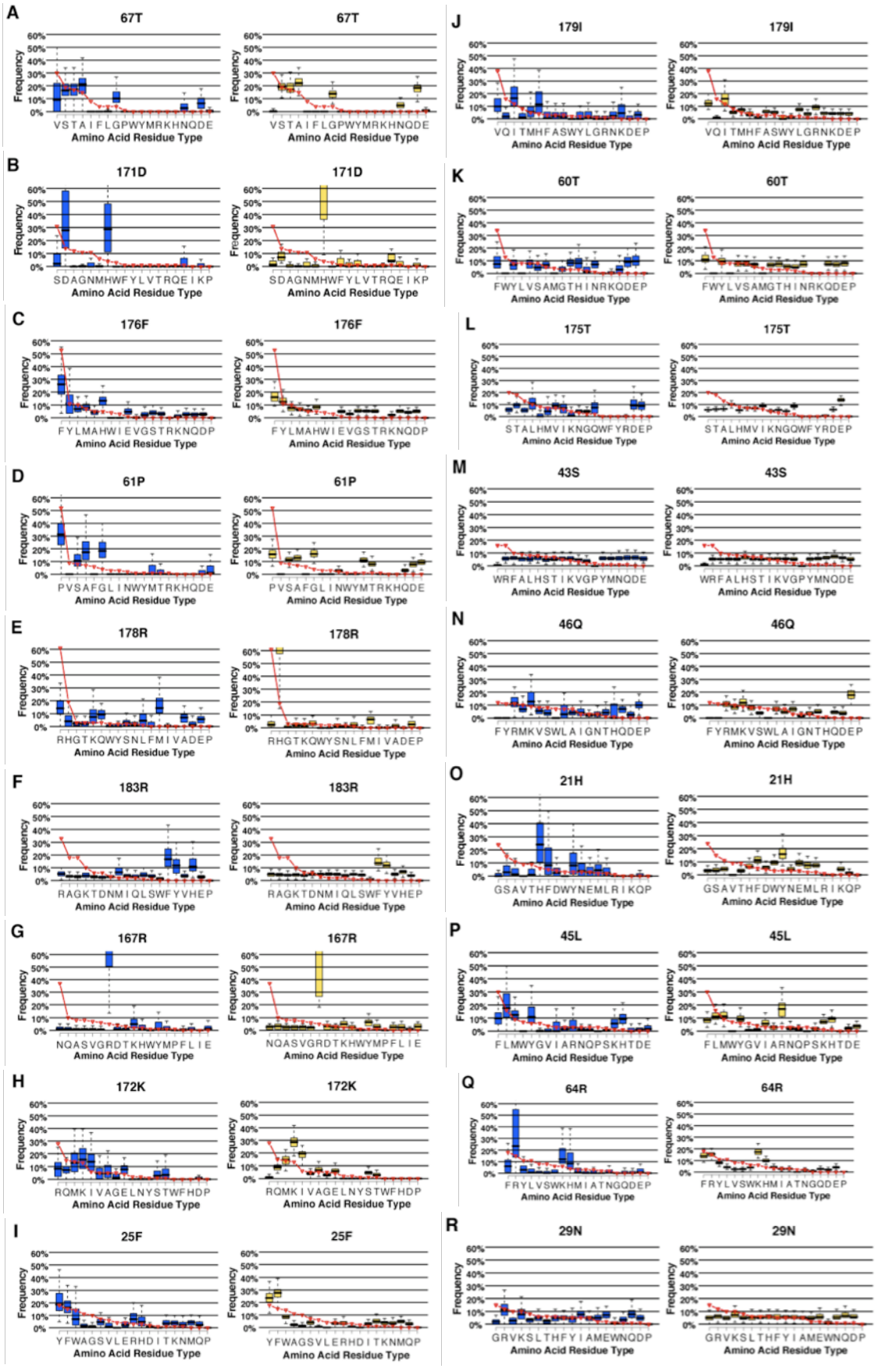
Variability of Fixed Backbone Tolerance Profiles at Four Interface Positions

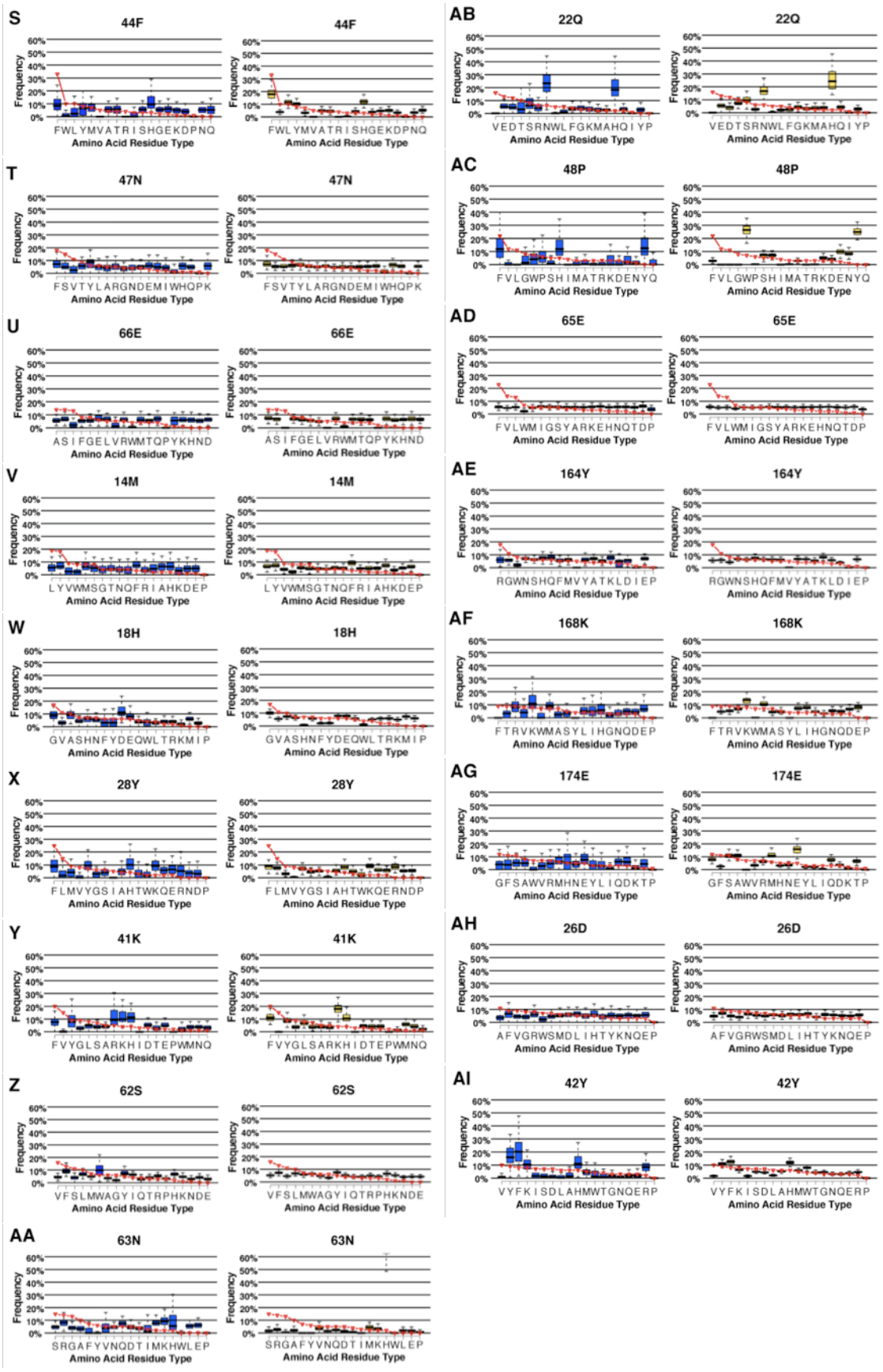
As described in methods, the score calculated for the optimized rotamer combinations for the native sequence on each backbone ("native score") was used throughout the remainder of the simulation as a threshold by which other sequences are accepted or rejected (Figure 3.1). Due to the stochastic nature of the Monte-Carlo optimization protocol, two different sets of low- scoring rotamer combinations appeared to be selected for the fixed backbone simulations at four positions: 171D, 167R, 63N, and 178R. In each case, this resulted in some fraction of the 100 independent runs on the fixed crystallographic backbone having a "native score" higher than that of the remainder of the runs, and thus accepting a different set of sequences. This caused a significant variation of the frequency of the most commonly observed residue type (for example variation for histidine in Figure 3.4F, and variation for arginine in Figure 3.4G, yellow bars). This problem seemed only to occur for polar residues and is hence likely to be linked to cases involving intricate alternative networks of polar interactions and complications in modeling these interactions accurately.

B.1.2. Supplementary Figures



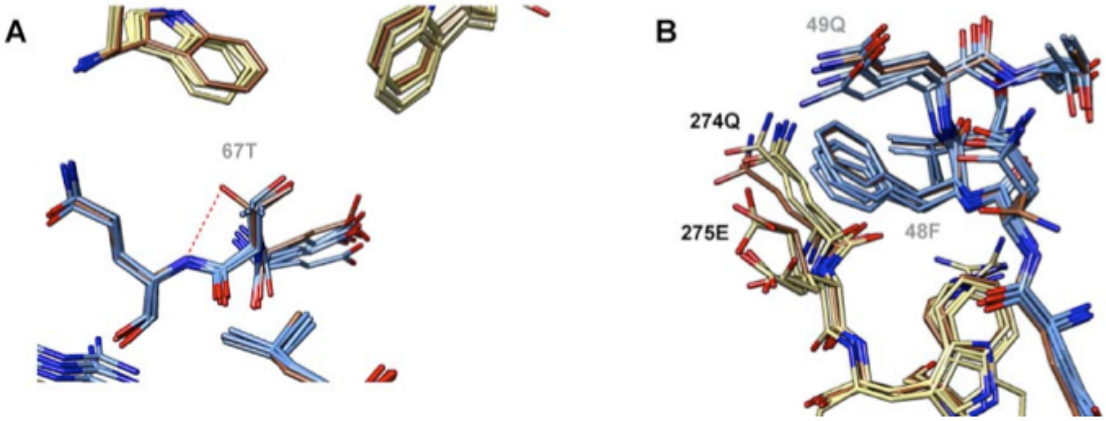
Supplementary Figure B.1: Structural Comparison of Crystallographic and Computational hGH-hGHR B-Values
(A) Color coded average per-residue B-values as taken from 1A22.pdb. Chimera was used to divide the histogram of average experimental B-values into groups of sizes mimicking the distribution plotted in (B). These groups are as follows: 9.5<white<15.99; 15.99<blue<23.26; 23.26<green<30.52; 30.52<yellow<37.79; 37.79<orange<45.05; red>45.05. (B) Color coded average per-residue B-values as calculated by GROMACS for the backrub generated ensemble and were colored onto the 100-member ensemble as follows: 0≤white<10; 10≤blue<20; 20≤green< 30; 30≤yellow<40; 40≤orange<50; red≥50. Note the cartoon radius depicted for each ensemble member in (B) has been significantly reduced compared to the sizing shown for 1A22.pdb in (A). All B-values calculated by GROMACS have been multiplied by a scaling factor of 10.


A
67T
67T
Frequency
60%
50%
40%
30%
20%
10%
0%
VSTAIFLGPWYMRKHNQDE
Amino Acid Residue Type
B
171D
171D
SDAGNMHWFYLVTRQEIKP
C
176F
176F
FYLMAHWIEVGSTRKNQDP
D
61P
61P
PVSAFGLINWYMTRKHQDE
E
178R
178R
RHGTKQWYSNLFMIVADEP
F
183R
183R
RAGKTDNMIQLSWFYVHEP
G
167R
167R
NQASVGRDTKHWYMPFLIE
H
172K
172K
RQMKIVAGELNYSTWFHDP
I
25F
25F
YFWAGSVLERHDITKNMQP
J
179I
179I
VQITMHFASWYLGRNKDEP
K
60T
60T
FWYLVSAMGTHINRKQDEP
L
175T
175T
STALHMVIKNGQWFYRDEP
M
43S
43S
WRFALHSTIKVGPYMNQDE
N
46Q
46Q
FYRMKVSWLAIGNTHQDEP
O
21H
21H
GSAVTHFDWYNEMLRIKQP
P
45L
45L
FLMWYGVIARNQPSKHTDE
Q
64R
64R
FRYLVSWKHMIATNGQDEP
R
29N
29N
GRVKSLTHFYIAMEWNQDP
Amino Acid Residue Type


S
44F
44F
T
47N
47N
U
66E
66E
V
14M
14M
W
18H
18H
X
28Y
28Y
Y
41K
41K
Z
62S
62S
AA
63N
63N
AB
22Q
22Q
AC
48P
48P
AD
65E
65E
AE
164Y
164Y
AF
168K
168K
AG
174E
174E
AH
26D
26D
AI
42Y
42Y
Frequency
Amino Acid Residue Type

Supplementary Figure B.2 Comparison of Computationally Generated Tolerance Profiles to Tolerance

Profiles of Experimentally Derived Sequences for all 35 hGH-hGHR Interface Positions Tolerance profiles generated by the flexible (blue box-plots) and fixed backbone (yellow box-plots) protocols are compared to the experimental tolerance profiles of folded sequences determined to bind hGHR (red lines) for all 35 hGH-hGHR interface positions. Interface positions are ordered as in Figure 3.3.

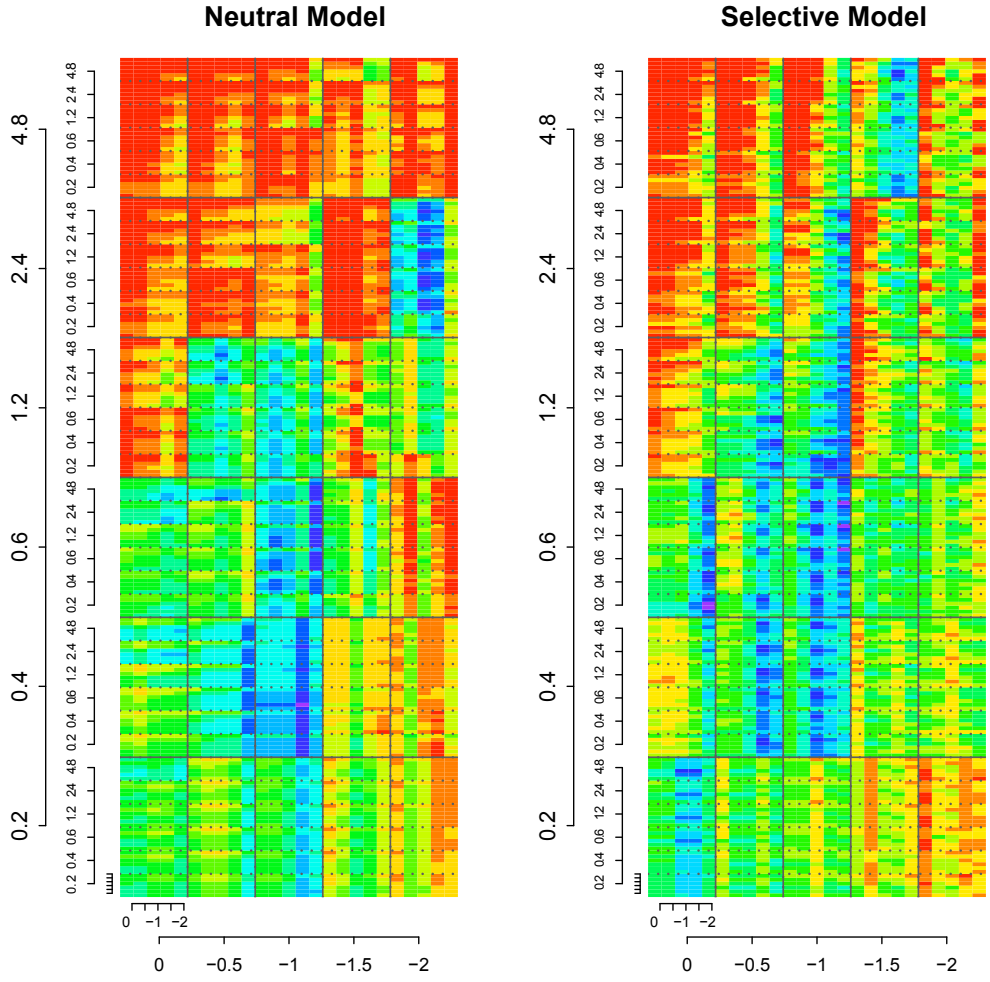


Supplementary Figure B.3 Structural Illustrations of the Possible Consequences of Backbone Flexibility

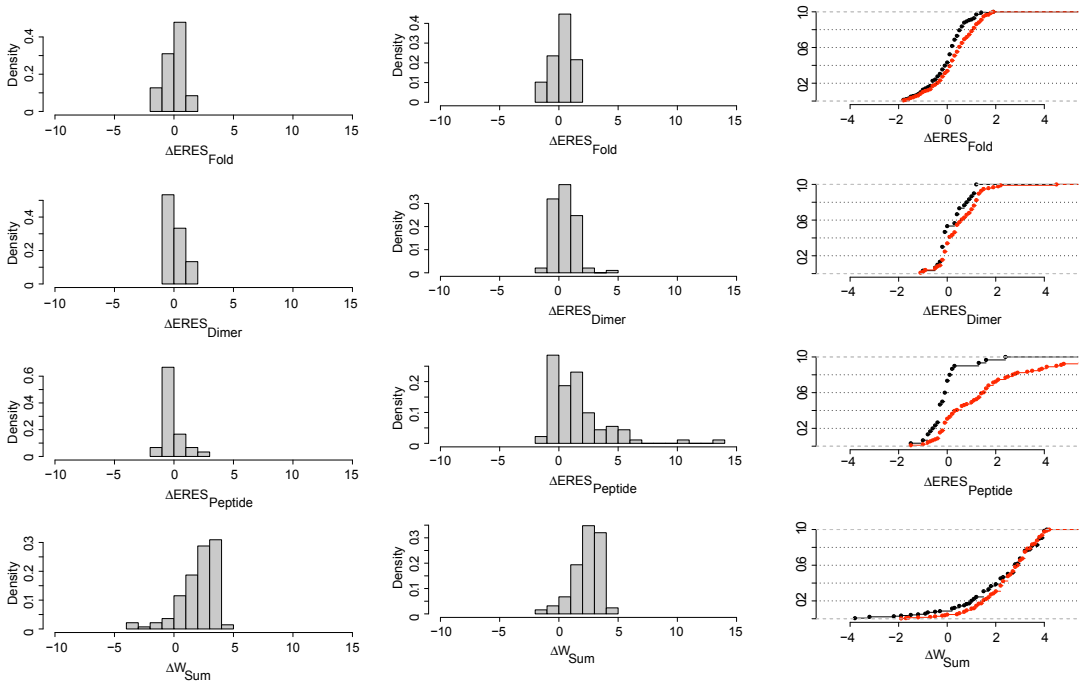
(A) Comparison of side chain placement observed in the minimized crystallographic hGH-hGHR complex (1A22.pdb,brown) to side-chain placement seen in three ensemble members (yellow, hGHR; blue, hGH) around residue 67T. The hydrogen bonding interaction between 67T and a backbone nitrogen seen in1A22.pdb (red dotted lines) are absent for each ensemble member. This results in an overall favorable prediction of replacement of 67T for 67V when predictions are made on the flexible backbone ensemble. (B) Comparison of side chain placement observed in the minimized crystallographic hGH-hGHR complex (1A22.pdb,brown) to side-chain placement seen in four ensemble members (yellow, hGHR; blue, hGH) after mutation of residue 48P to 48F. For the crystallographic structure, mutation to 48Y and 48W are predicted to be highly favored over the native 48P. However, predictions made using the flexible backbone ensemble strongly reduce substitution by W and somewhat reduce substitution by Y. This could be due to side chain rearrangements of 49Q and 213Q resulting in a somewhat smaller pocket surrounding residue 48, as suggested by the backbone shown in (B).

Appendix C. Chapter 4, Supplementary Materials

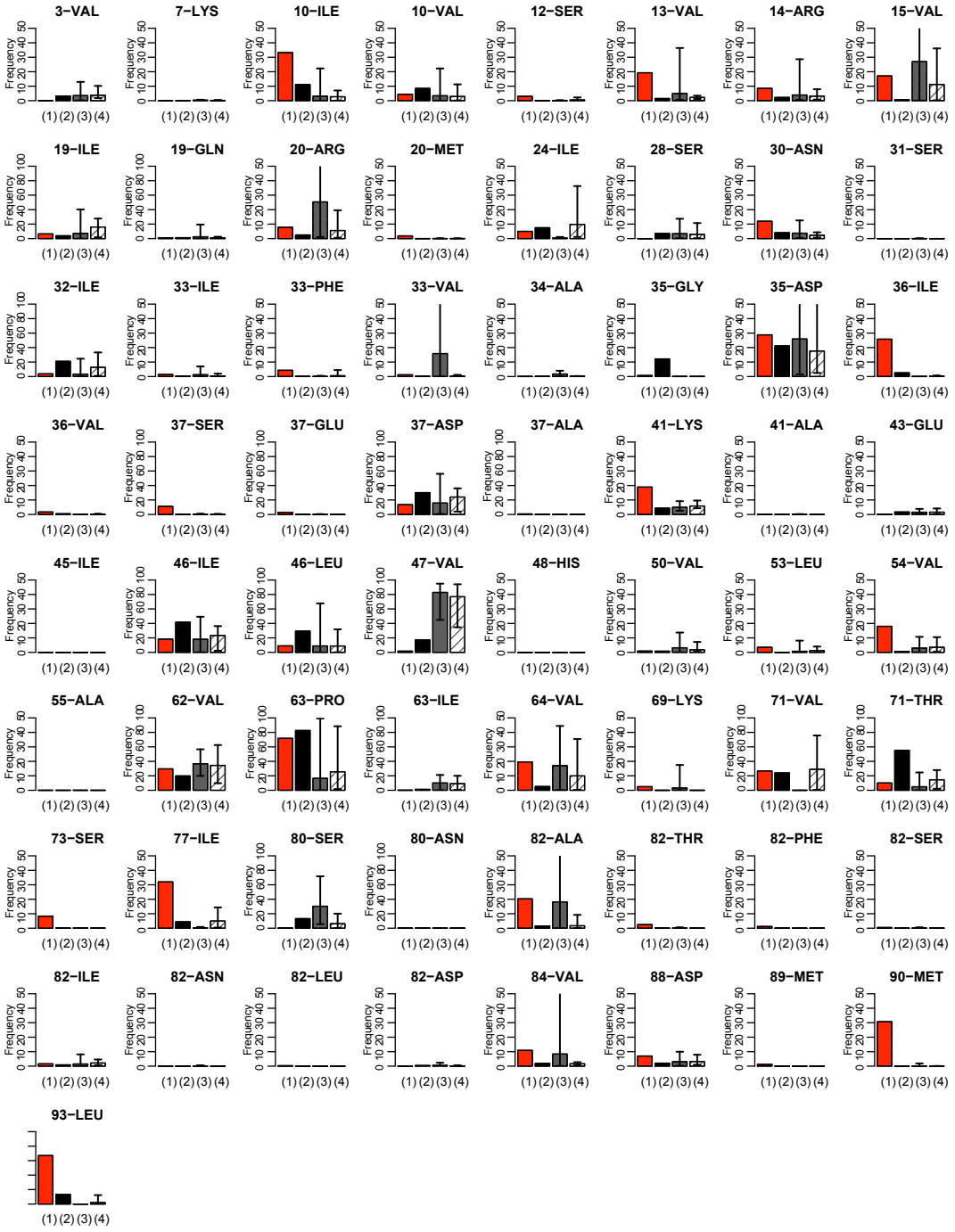
C.1.1. Supplementary Figures



Supplementary Figure C.1 Model performance as parameters are varied.
Variation of parameters for neutral **(A)** and selective pressure **(B)** computational models is shown. Fold (outermost labels, y-axis), dimer (middle labels, y-axis) and peptide (inner labels, y-axis) parameters weights (see Eq. 1, main text) were varied systematically along with score value added to native hydrophobic (outer labels, x-axis) or native polar (inner labels, x-axis). For each parameter combination the number of protein residue sties correctly predicted as having low (1-5%), medium (5-20%) or high (>20%) mutation rates under both a neutral (no protease inhibitors) and selective (1-9 protease inhibitor) models was recorded. The correct recovery (averaged over the three bins) of each parameter set is color-coded, with blue depicting strong discrimination among sites of low, medium and high mutation rate and red indicating poor discrimination. Parameter sets selected for model (see Supplementary Table C.3) are found within the black boxes on each plot. Variation of parameter adjusting for disfavoring polar to hydrophobic substitutions is not shown.

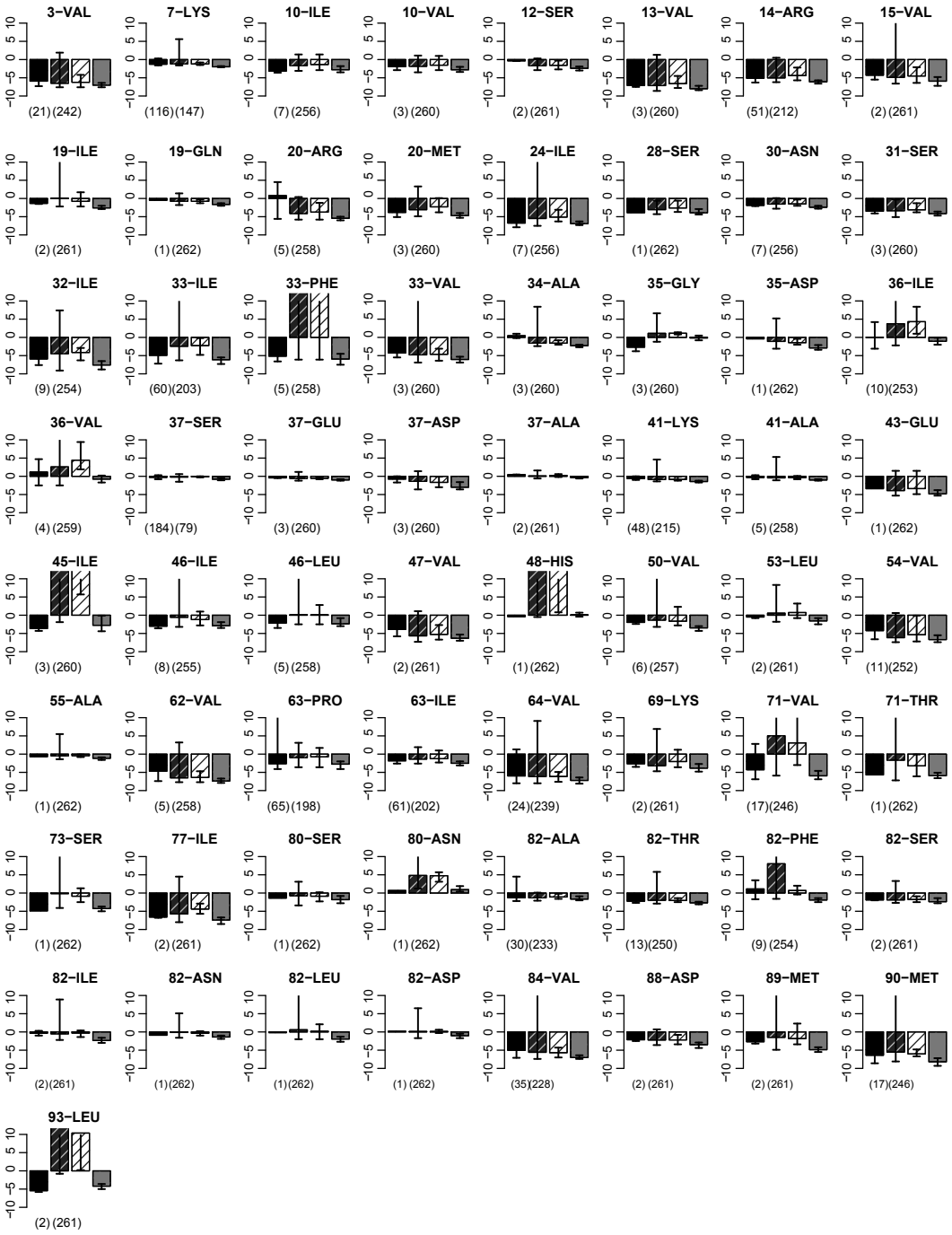


Supplementary Figure C.2 Distribution of ERES values for tolerated amino acid types.
A histogram showing the distribution of change in $ERES_{FOLD}$ (top row) $ERES_{DIMER}$ (2nd row) and $ERES_{PEPTIDE}$ (3rd row) relative to the native ERES scores is shown for mutations computationally predicted to be tolerated by the neutral **(A)** and selective **(B)** models. The distribution of W_{SUM} values for tolerated mutations relative to the native W_{SUM} value is also shown for each model (bottom row). Cumulative distribution functions calculated for the histograms shown for the neutral (black lines) and selective (red lines) models are given in **(C)**.



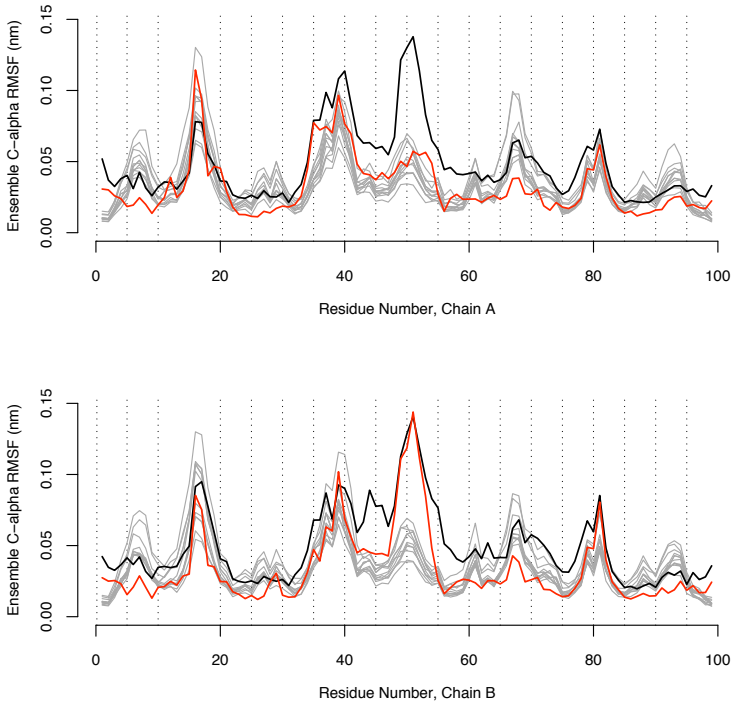
Supplementary Figure C.3 Comparison of predicted frequencies with and without backbone flexibility for all crystallographic mutations.

Bar plots depicting the frequency with which each mutation was observed in the Stanford database after protease treatment ((1), red bars), predicted by the full ensemble model ((2), black bars), predicted by 11 fixed backbone structures crystallized in the absence of mutation (3), grey bars), or predicted by 11 "backrub" ensembles generated from each fixed backbone structure in (3) ((4), striped bars) are shown. The 67 mutations shown represent all the mutations contained within at least one crystallographic structure.



Supplementary Figure C.4 Comparison of $ERES_{FOLD}$ values with and without backbone flexibility for all crystallographic mutations.

Bar plots depicting the distribution of $ERES_{FOLD}$ values calculated on all crystal structures containing each mutation at the time of crystallization (black bars), all crystal structures not containing each mutation at the time of crystallization (but possibly containing other mutations; black bars with white stripes), 11 crystal structures crystallized in the absence of any mutation (white bars with grey stripes), or 11 "backrub" ensembles generated from each fixed backbone structure (grey bars) are shown. The 67 mutations shown represent all the mutations contained within at least one crystallographic structure. The number of structures containing each mutation is shown in parenthesis under each black bar while the number of structures without each mutation is given in parenthesis under each black bar with white stripes.



Supplementary Figure C.5 Comparison of RMSF calculated for crystallographic and computational ("backrub") structural ensembles

C-alpha RMS fluctuations for the 263 crystallographic structures used to calculate $ERES_{FOLD}$ and $ERES_{DIMER}$ scores and the 16 crystallographic structures used to calculate $ERES_{PEPTIDE}$ scores are shown in black and red respectively. For comparison, C-alpha RMS fluctuations for ensembles of structures independently generated by using the backrub protocol (as described in **Section 4.3.4**) starting from one of 11 crystallographic structures with the native subtype-B consensus sequences are shown in grey. RMS fluctuations are similar among all ensembles for most of the 99 residues in each HIV-protease chain. Note, the crystallographic structures show large variation in the flap region (near residue 50) for both chains, as some crystallographic conformations have been solved in the "flap open" conformation. The peptide bound structures show an asymmetric behavior between the two chains for this region while the backrub structures (all generated from a "flap-closed" starting conformation) show smaller fluctuations in the flap dynamics for both chains. The positions of residue 1 and residue 99 are fixed during the entire backrub protocol, and thus show RMS fluctuations of zero.

C.1.2. Supplementary Tables

1A30	1A8G	1A8K	1A94	1A9M	1AAQ	1AID	1AJV	1AJX	1AXA
1BDL	1BDQ	1BDR	1BV7	1BV9	1BWA	1BWB	1C6X	1C6Y	1C6Z
1C70	1D4H	1D4I	1D4J	1D4S	1D4Y	1DIF	1EBW	1EBZ	1EC0
1EC1	1EC2	1EC3	1F7A	1FQX	1G2K	1G35	1G6L	1GNM	1GNN
1GNO	1HBV	1HIH	1HIV	1HOS	1HPO	1HPS	1HPV	1HPX	1HSG
1HTE	1HTF	1HTG	1HVH	1HVI	1HVJ	1HVK	1HVL	1HVR	1HVS
1HWR	1HXB	1HXW	1IIQ	1IZH	1IZI	1K1T	1K1U	1K2B	1K2C
1K6C	1K6P	1K6T	1K6V	1KJ4	1KJ7	1KJF	1KJG	1KJH	1KZK
1LV1	1LZQ	1M0B	1MER	1MES	1MET	1MEU	1MRW	1MRX	1MSM
1MSN	1MT7	1MT8	1MT9	1MTB	1MUI	1N49	1NH0	1NPA	1NPV
1NPW	1ODW	1ODX	1ODY	1OHR	1PRO	1QBR	1QBT	1QBU	1RL8
1RPI	1RQ9	1RV7	1SBG	1SDT	1SDU	1SDV	1SGU	1SH9	1SP5
1T3R	1T7J	1T7K	1TCX	1TSQ	1TSU	1TW7	1U8G	1VIJ	1VIK
1W5V	1W5W	1W5X	1W5Y	1WBK	1WBM	1XL2	1XL5	1YT9	1YTG
1YTH	1ZLF	1ZTZ	2A1E	2A4F	2AID	2AOC	2AOD	2AOE	2AOF
2AOG	2AOH	2AOI	2AOJ	2AQU	2AVM	2AVO	2AVQ	2AVS	2AVV
2AZC	2B60	2B7Z	2BB9	2BBB	2BPV	2BPW	2BPX	2BPY	2BPZ
2BQV	2CEJ	2CEM	2CEN	2F3K	2F80	2F81	2F8G	2FDD	2FDE
2FGU	2FGV	2FLE	2FNS	2FNT	2FXD	2FXE	2HB3	2HC0	2HS1
2HS2	2I0A	2I0D	2I4D	2I4U	2I4V	2I4W	2I4X	2IDW	2IEN
2IEO	2NMW	2NMY	2NMZ	2NNK	2NNP	2NPH	2NXD	2NXL	2NXM
2O4K	2O4L	2O4N	2O4P	2O4S	2PK5	2PK6	2PQZ	2PSU	2PSV
2PWC	2PWR	2PYM	2PYN	2Q3K	2Q54	2Q55	2Q5K	2Q63	2Q64
2QCI	2QD6	2QD7	2QD8	2QHY	2QHZ	2QI0	2QI1	2QI3	2QI4
2QI5	2QI6	2QI7	2QNN	2QNP	2QNQ	2R5P	2R5Q	2UPJ	2UXZ
2UY0	2Z4O	3AID	3B7V	3B80	3BVA	3BVB	4PHV	5HVP	7UPJ
9HVP									

Supplementary Table C.1 Structural ensemble PDB codes for fold and dimer stability calculations. The table list each of the 263 pdb codes used for determining $ERES_{FOLD}$ and $ERES_{DIMER}$ scores.

Constraint	Structure	Substrate	Sequence
1	1MT7.pdb	Matrix-Capsid	VSQNY-PIVQN
	1KJ4.pdb	Matrix-Capsid	VSQNY-PIVQN
2	1MT8.pdb	Capsid-p2	KARVL-AEAMS
	1F7A.pdb	Capsid-p2	KARVL-AEAMS
3	1MT9.pdb	p1-p6	RPGNF-LQSRP
	1KJF.pdb	p1-p6	RPGNF-LQSRP
4	1TSU.pdb	Nucleocapsid-p1	ERQAN-FLGKI
	1TSQ.pdb	Nucleocapsid-p1	RQVN-FLGKIN
	2FNS.pdb	Nucleocapsid-p1	RQAN-FLGKIN
	2FNT.pdb	Nucleocapsid-p1	RQVN-FLGKIN
5	1KJ7.pdb	Nucleocapsid-p2	KATIM-MQRGN
6	1KJG.pdb	RT-RNAse H	GAETF-YVDGA
	2NXD.pdb	RT-RNAse H	GADIF-YLDGA
	2NXL.pdb	RT-RNAse H	GAEVF-YVDGA
	2NXM.pdb	RT-RNAse H	GAQTF-YVDGA
7	1KJH.pdb	RNAse H-IN	IRKIL-FLDGI
8	Modeled	TF-PR	VSFNF-PQITL
9	Modeled	AutoP	PQITL-WQRPL
10	Modeled	PR-RT	CTLNF-PISPI

Supplementary Table C.2 Structural ensemble of PDB codes and peptides used for substrate calculations.
For each of the 10 endogenous peptides considered, the PDB codes of all crystal structures used, as well as their peptide sequence present in the crystallographic structure, are given. All peptides are denoted from P5-P5' except for 1TSQ, 2FNS, and 2FNT which are given from P4-P6'. Amino acids not visible within the X-ray crystallographic density, and thus not present in computational simulations, are shown in grey. Amino acids colored red are peptide mutations observed in response to the HIV-1 protease drug resistance mutation V82A. Blue amino acids were computationally engineered for tighter protease binding affinity.

Weighted Term	Neutral Model	Selective Model	Parameter Set
W_{FOLD}	0.4	0.6	0.2 0.4 0.6 1.2 2.4 4.8
W_{DIMER}	0.6	2.4	
$W_{PEPTIDE}$	0.6	4.8	
$FAVOR_{NATIVE}$	1	1	0 -0.5 -1 -1.5 -2
$FAVOR_{POLAR\_NATIVE}$	1.5	1	
$PENALTY_{POLAR \rightarrow HP}$	1.5	1.5	

Supplementary Table C.3 Parameter values used in the computational model.
All parameter values used in the neutral and selective models are as given. For parameter optimization, all possible combinations of parameter values listed in the parameter set were tested for their predictive ability in determining overall mutational frequencies at each of the 99 sequence sites in HIV-1 protease (see Supplementary Figure C.1).

	A	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A		1	1		1							1			1	1	1		
D	1		1		1	1					1						1		1
E	1	1			1			1					1				1		
F							1		1						1		1		1
G	1	1	1											1	1		1	1	
H		1							1		1	1	1	1					1
I				1				1	1	1				1	1	1	1		
K			1				1			1	1		1	1		1			
L1				1		1	1			1		1	1	1			1		
L2				1			1			1					1		1	1	
M							1	1	1					1		1	1		
N		1				1	1	1							1	1			1
P	1					1			1				1	1	1	1			
Q			1			1		1	1			1		1					
R1					1	1			1			1	1		1			1	
R2					1		1	1		1					1	1		1	
S1	1			1					1			1				1		1	1
S2					1		1					1		1		1			
T	1						1	1		1	1	1		1	1				
V	1	1	1	1	1		1		1	1									
W					1				1					1	1				
Y		1		1		1					1				1				

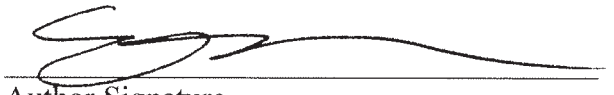
Supplementary Table C.4 MUT_{PROB} values used in the computational model.
The table shows MUT_{PROB} values for every possible HIV-1 protease mutation. Starting from a given native amino acid type (horizontal rows) MUT_{PROB} values to all other amino acid types (vertical columns) are shown. White boxes denote MUT_{PROB} values of zero. All boxes with a value of one were considered as tolerated for the naive single nucleotide mutational model, while all boxes colored blue were considered to be tolerated for a null model of chemically similar amino acid types. Leucines and arginines were assigned to one of two codons as follows: L1, residues (5,10,19,23,63,76,89) or L2, residues (24,33,38,90,97); R1, residue (8) or R2, residues (41,57,87). There were no serines in the HIV-1 protease sequence to assign to S1 or S2.

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

Please sign the following statement:

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.



Author Signature

10/26/2009

Date